Defense Centers of Excellence
for Psychological Health and
Traumatic Brain Injury

PROGRAM EVALUATION GUIDE

MODULE 7

Analyzing Quantitative Data

May 2017 | 2nd Edition

dcoe.mil

Table of Contents

Overview of the Program Evaluation Guide

This Program Evaluation Guide (PEG) is developed and published by the Defense Centers of Excellence for Psychological Health and Traumatic Brain Injury (DCoE). Program evaluation is an important part of the DCoE **mission** and helps military program administrators and leadership assess and improve service quality and outcomes. By making program evaluation an inherent part of everyday program activities, we create a culture of effectiveness to better build a sustainable, efficient and well-integrated continuum of prevention and care services for military members, their families and veterans.

The first edition of the PEG, published in July 2012, provided a standardized approach to program evaluation for psychological health and traumatic brain injury (TBI) program leaders. This version of the PEG (2nd Edition) has been updated and revised to reflect the most current needs of psychological health and TBI programs. This edition of the PEG is organized as a series of modules containing content specifically designed for use by program administrators or other staff members tasked with internal program evaluations as part of their duties within Defense Department psychological health and TBI programs. This PEG is designed for those who have limited prior knowledge and experience with the conduct of program evaluation activities.

Purpose and Use of the PEG

This PEG is one part of a collection of trainings, toolkits and support services offered by DCoE to assist personnel at the program level in developing their capabilities to conduct internal program evaluation activities. The PEG is designed for use in coordination with other training materials, such as DCoE's program evaluation and improvement webinar series, references provided in the PEG and webinar series, consultation with experts and other resources that may be available to program personnel.

The modules in this PEG are not intended to serve as a substitute for formal coursework on evaluation methods, statistics or data management. In addition, because the PEG is intended for use by a wide variety of programs, it will not provide specific guidance to programs on best practices for clinical or non-clinical services. Finally, the PEG is not intended as a manual for how evaluators who are external to a program should conduct their activities. However, the information herein will generally be useful in helping program personnel become more familiar with the evaluation process and consequently more effective in responding to external evaluation initiatives.

Analyzing Quantitative Data

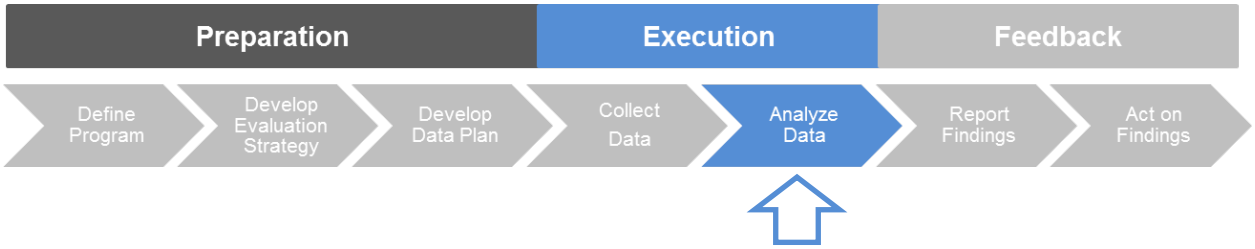
Purpose and Use of this Module

Once data are collected, organized and stored, the program is ready to move to the next step of the evaluation process, analyzing data.

This module, "Analyzing Quantitative Data," is one of three PEG modules focused on different aspects of analyzing data. It is specifically designed to assist program personnel in their efforts to analyze and interpret data derived from quantitative evaluation methods. Analysis of qualitative data and cost-related quantitative data are the focus of other modules.

In this module, we provide an overview of the major types of quantitative data and statistical methods commonly used in program evaluation, followed by approaches used to summarize and compare results of quantitative analyses. In addition, this module provides some insight on methods that can be used to enhance the validity and reliability of results.

Because evaluation efforts will differ across every program, this module provides broadly applicable guidance on procedures used to analyze quantitative data as part of a program evaluation effort.



Overview of Quantitative Data Analysis

Stephen Few, an author and business analyst, said, "Numbers have an important story to tell. They rely on you to give them a clear and convincing voice" (Brent, 2016). The focus of this module is about making sense of the data we collect and learning what they reveal about the program. It is about what to do with data once they have been collected. The aim of data analysis and interpretation is to condense large amounts of information into actionable form and communicate important findings. Data analysis should provide enough detail to be informative, but not so much that the program stakeholders cannot understand and absorb it.

Quantitative and qualitative analysis strategies are alike in that they seek to organize and reduce a large amount of fine-grained information into a relatively small number of patterns that describe or explain. As such, both analysis strategies involve transforming data from its raw form into something more readily understandable in relation to the purpose of a program evaluation effort. When used in program evaluation, both quantitative and qualitative analysis strategies also have the same end goal: interpreting information about a program to learn something about how it operates and how it affects its participants.

Quantitative methods are ideally suited for identifying the who, what, where and when among target populations. In addition, quantitative data analysis helps evaluators to classify information, make comparisons, count relevant elements and even construct more complex

statistical models in an attempt to explain what is observed. Quantitative findings can often be generalized to a larger population provided appropriate statistical techniques are used along with samples representative of the target population. Thus, quantitative analysis, when conducted well, can provide insights that go beyond a single program site.

There are two broad types of quantitative data analysis methods: descriptive and inferential statistics. The sections that follow discuss how these methods are used to address common process and outcome questions of interest to evaluators and program administrators alike.

Descriptive Statistics

In program evaluation, **descriptive statistics** are used to show the recurrence of certain responses or characteristics, what is average or typical for a group of individuals, and how spread out participant data are on variables of interest. Descriptive statistics include a variety of techniques such as counts (i.e., frequencies and percentages), measures of central tendency (i.e., mean, median and mode), and measures of variability (i.e., range, variance and standard deviation). Descriptive statistics are commonly used to summarize information and may be featured in tables, charts and graphs that illustrate key characteristics of the data or demonstrate how variables relate to one another.

Counts (frequencies and percentages) are used to describe how common a phenomenon or characteristic is within a sample or population. One common way to summarize frequency data is a cross-tabulation table. These tables consist of rows displaying values for one variable of interest and columns displaying values for observations. Cross-tabulation tables can compare several groups or time periods at once. You can use these tables to illustrate any of the statistical methods discussed above.

Table 1 provides an example of a cross-tabulation table. In this example, the program information displays changes in frequency of headaches that may relate to the impact of program participation in a stress reduction program targeting several measurable health outcomes (e.g., headaches, blood pressure). As indicated, the first column contains labels for two categories of headache frequency. Additional columns show the number of participants out of a sample of 1000 who provided survey data on headache frequency before and after program participation and again at a six-month follow-up. Alternatively, these numbers could be presented as percentage values and/or placed in a bar graph to visually display the pattern of decreasing headache frequency from baseline to post-intervention, which are mostly maintained during the six-month follow-up period.

Table 1: Example of Headache Frequency Cross-tabulation at Baseline, Post-intervention, and Six-month Follow-up for Participants in a Stress Reduction Program

Headache Frequency	Baseline	Post-Intervention	6 Months Post-Intervention
Headaches Present > 1x/week	345	192	213
Headaches Present ≤ 1x/week	655	808	787

More complex cross-tabulation tables can also contain detailed information that could show impact across varying demographic characteristics such as age, military rank or gender. Overall, the goal of cross-tabulation tables and graphs representing counts is to take the raw

data collected and present it in a way that is easy to see and understand quickly and effectively, thus demonstrating the essence of the relationship between variables of interest and observed data.

Measures of central tendency (mean, median and mode) allow us to summarize the average response for an entire data set with a single value. For example, the blue-shaded region in Figure 1 below depicts the measures of central tendency for the information listed in Table 1 above (converted to monthly frequencies). The **mode** (i.e., most frequent response) indicates the majority of participants experienced fewer headaches per month resulting in a positively skewed graph. This results in the tail of the blue graph stretching to the right where the **mean** (i.e., statistical average of the responses) and **median** (i.e., middle number in the distribution of responses) would be pulled to the right. If the data were negatively skewed, meaning that the tail in the pink-shaded region of the graph stretched to the left, the median and mean would be pulled to the left as shown below. This would indicate the majority of participants experienced a higher number of headaches per month.

The center, gray-shaded region of Figure 1 depicts a graph for data that is normally distributed, meaning that it is bell-shaped with the mean and median aligned in the center of the curve. For the example in Table 1 to have been normally distributed, the number of headaches per month would have been more evenly distributed around a central frequency of headaches experienced by the majority of participants.

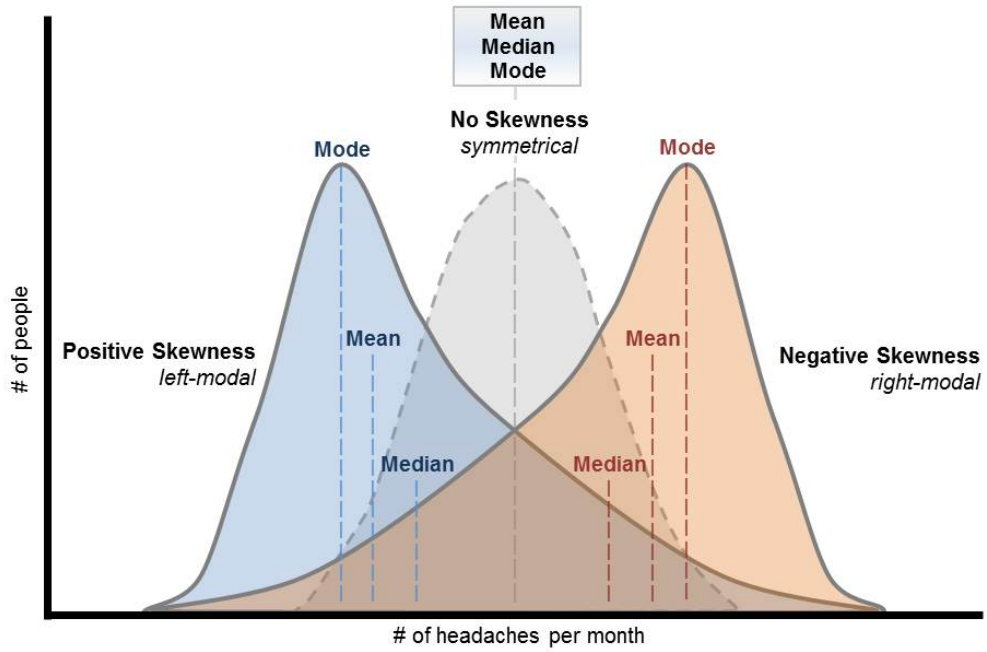


Figure 1: Example of Central Tendency Measures for Headache Frequency

Whether data are normally distributed has important implications for the types of analyses used. Distributions that show high levels of skew (i.e., stretched out to the left or right) or kurtosis (i.e., pointy vs. rounded shape due to a narrow range of responses) require

specialized techniques or transformations that may require consultation with an expert in statistics.

Measures of variability (range, variance and standard deviation) tell us how much observations in a data set differ from one another. **Range** is the size of the interval between the largest and smallest values. For example, if the highest number of headaches per week in a sample was 30 and the lowest reported number was 0, then the range would be 30. **Variance** measures how far a set of data points are spread out from the mean using squared units. In order to standardize these units, **standard deviation** is calculated as the square root of variance, which converts the amount of variation or dispersion back into the same units as the original set of data values. In normal distributions, approximately 68 percent of participants' responses fall within plus or minus one standard deviation unit, while approximately 95 percent fall within two units. A low standard deviation indicates the data points tend to be close to the mean, while a high standard deviation indicates the data points are spread out over a wider range of values. Visually, distributions with low standard deviation have a peaked or pointy shape, while high standard deviation looks more flattened. Variability has important implications for determining the statistical significance of results when conducting inferential analysis, which is discussed in the following section.

In addition to their ability to describe the general features of data, descriptive statistics can also assist with the inspection of data for:

- Determining the most appropriate types of inferential analysis for the evaluation questions
- Identifying data entry errors, incomplete data or outliers
- Determining whether statistical assumptions are met (e.g., normal distribution, minimum sample size)

Inferential Statistics

Inferential statistics are used to test hypotheses or examine evaluation questions by drawing conclusions about a population based on data collected from a smaller sample. They are used to estimate whether the measured differences between two or more groups on variables of interest are substantial enough to expect similar differences in the population. For example, studies of a treatment might want to know whether Group A that received a treatment has better outcomes than Group B that did not. Some techniques include the use of effect sizes, confidence intervals, regression analysis and group comparisons. Figure 2 below provides an overview followed by definitions. Although this section describes inferential statistics, more specialized evaluation strategies may be needed. Inferential techniques use statistics to draw inferences about a population based on information from a sample of individuals from that population. Additional resources are listed in the "References" and "Selected Resources for Additional Study" sections at the end of this module.

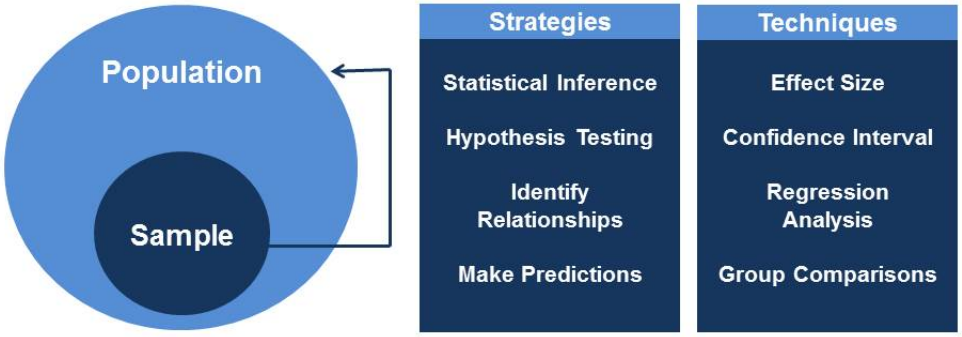


Figure 2: Inferential Statistics Overview

Common inferential statistics are described below.

Effect Size: An effect size is a quantitative measure of the strength and direction of a difference between the average score of participants either before and after an event (e.g., treatment), or between two different groups (e.g., a treatment and comparison group) (Ellis, 2010). For example, Cohen's *d* is a common effect size measure that uses standard deviation units. After administering and analyzing pre- and post-test data, if knowledge increased an average of one-half standard deviation, *d* would equal 0.5.

Confidence Interval: Programs can use confidence intervals to assess the probability that their sample statistics are a true representation of the target population, e.g., the higher the probability, the higher the level of confidence. A confidence interval is a range of values likely to include the population mean. It is calculated by using the standard error of the mean to estimate the variability between multiple sample means taken from the same population. A 95 percent confidence interval around an effect size could be d = 0.50 (95% CI = .25 - .75). Since the confidence interval around the value *d* = 0.50 does not include 0.00, it would be considered significant in this case. For example, a program may want to compute a 95 percent confidence interval for the number of program participants with suicide attempts six months after program completion as part of process evaluation and risk management (Brown, 2016). By understanding how likely a given risk is to occur, the program can manage the risks of a non-occurrence accordingly (Richards, 2016).

Regression Analysis: Regression analysis attempts to determine the strength of the relationship between one criterion variable and a series of other predictor variables. For example, regression could be used to predict the likelihood of developing depression, given a collection of risk factors such as family history, exposure to stress and alcohol use.

Group Comparisons: The techniques compare average values from two or more different groups or populations to determine whether they are statistically significant. For example, *t*-tests and analysis of variance (ANOVA) are commonly used to determine whether differences between groups are greater than one would expect given the differences, or variation, within groups. For example, a program may use group comparison to examine a participant group that receives medication and therapy to a group that receives medication only and depression level post program completion to assess the impact of the interventions

on depression. The comparison group design most likely to produce valid results is when both groups have similar characteristics in the areas of demographics, geographic location, data collection consistency and participant motivation levels (Coalition for Evidence-Based Policy, 2014).

The overall purpose of using these inferential statistics techniques is to assess the probability that the results are real rather than due to chance. The analysis and determination of significance may depend on a number of factors, including industry standards, similar research studies, comparison groups and/or methods used for inferential analysis, such as *t*-tests.

Process Analysis

Quantitative data can be used to answer a variety of evaluation questions focused on program processes. Process analyses make use of quantitative data by reflecting program inputs, activities and outputs, which may be organized in a program logic model as described in previous modules:

Inputs: the facilities, equipment, staffing and financial resources required to operate a program

Activities: clinical, outreach or educational activities offered by the program

Outputs: participation numbers and demographics, as well as measurable products that result from program activities

With these data, programs can address a number of important questions related to program processes such as the questions below, which are often of interest to program managers and evaluators.

Common Process Questions

Fidelity: ***Were the program's activities or services implemented as intended or in accordance with a set of prescribed practices?*** Programs should generally be designed on the basis of scientific evidence and use a set of structured activities embedded in procedural documents (e.g., standard operating procedures or a treatment manual) that intend to produce changes in program participants. Thus, it is important to ensure the program is implemented with the same evidence-based processes that have been shown to work in other settings. To examine fidelity, quantitative data are used to compare intended activities and outputs as illustrated through consideration of coverage, as well as content, frequency and duration. Likewise, for large-scale programs, different sites may be compared to one another on indicators of fidelity to identify areas for improvement.

- **Coverage** examines whether the program is reaching its intended participant population by comparing data on the number and type of people the program has actually served with that which the program set out to serve. In some cases, a program may be serving too many or too few participants, which has important consequences for staffing and budgets. In other cases, the population served may be different than the one for which the program was set up.

- **Content, frequency and duration** measures provide concrete indicators for how well program activities match the original design of program services. For example, if clinicians are not able to provide services at the frequency designed or are not

covering core content, then a treatment may be less effective than it could be, and further investigation can reveal what is needed to enhance fidelity to the original model.

Satisfaction: ***How do stakeholders perceive the quality of the program?*** Quality concerns are of pivotal importance to programs, and data from stakeholder groups, such as participants, staff, leadership and decision makers, can be examined to determine how well a program meets their needs. Every program needs an effective process for tracking satisfaction and feedback from these stakeholders. Satisfaction ratings can be tracked using quantitative measures such as questionnaires or evaluation cards, and changes over time can be analyzed to assess quality improvement efforts.

The purpose of asking such questions is not only to analyze program processes but also to answer the question "How do processes affect program outcomes?" Answering questions about how well programs maintained fidelity or how satisfied participants were provides useful information for understanding how a program produces change among its participants, or perhaps why it did not produce as much change as desired. Without such process information, it is impossible to know why a program's results were as they were and how to refine a program to maximize benefits to participants. Below are examples of how these concepts can be applied to a program.

Example: Process Evaluation

The following figures and tables provide examples of process analyses for a hypothetical program, Program Sierra, including metrics for the target population's demographic characteristics, participant satisfaction ratings and the frequency of program activities.

Mission: At Program Sierra, we seek to ensure that service members who are wounded, ill or injured successfully reintegrate into civilian life or return to duty in the military. By performing our mission effectively, we hope to enhance force readiness and improve the quality and efficiency of services across the Defense Department.

Key evaluation questions

- Was Program Sierra implemented with fidelity?
  - What are the demographic characteristics of Program Sierra's participants?
  - How much of the target population is the program reaching?
- How satisfied are the participants with Program Sierra's services?

Program Sierra's evaluation team, led by its program manager, decided to conduct a process evaluation to determine whether the program was implemented with fidelity. The evaluation team began with the Process Analysis Matrix, provided in Appendix A, as a guide to focus their initial efforts on the important components of fidelity. The tables and graphs below are part of the process measures and quantitative analyses that were conducted as part of Program Sierra's process evaluation.

When comparing the number of individuals who participated in Program Sierra to those that

were targeted, Table 2 shows the program is reaching 87 percent of the targeted population as a whole. Subgroup analyses showed that populations were represented equally well across service components. However, when considering participant sex, only 70 percent of females are being reached by the program while 93 percent of males were reached. This would indicate the program may need to make more effort to engage female service members. Further analyses may be performed for additional elements of fidelity (i.e., content, frequency and duration).

Table 2: Coverage and Participant Demographics

	Target Population	Participant Population	
Demographics	Number (*N*)	Number (*n*)	Percent (%)
Total	29,694	25,931	87
Service Component			
Active Duty	22,959	19,915	87
National Guard	3,481	3,251	93
Reserve	3,254	2,765	85
Participant Sex			
Males	22,271	20,744	93
Females	7,423	5,187	70

Program is reaching most of the target population

Program is reaching a larger proportion of males (93%) than females (70%)

Program Sierra's evaluation team also distributed a questionnaire among program participants that asked "How satisfied are you with the services offered?" and measured responses on a five-point scale as summarized in Table 3.

Interpretation of the percentages may be enhanced by combining response options. For example, combining the percentages for the Extremely and Very Satisfied response options tells us that 45 percent of respondents were satisfied with their experience. Likewise, combining the percentages for the remaining response options indicates that 50 percent of the program participants were Not At All, Somewhat, or Neither Satisfied nor Dissatisfied with the services. Note that 5 percent of the participants did not respond to the question. Thus, there is room for improvement, and qualitative data or data from more specific quantitative items (e.g., satisfaction related to program staff, facilities or activities) could provide valuable information about how program services could be improved.

With half of the participants not being satisfied, the program manager may need to explore a possible relationship between low satisfaction and poor program outcomes, which illustrates just how process analyses can impact outcome analyses. Further, qualitative data could be used to provide additional explanatory information about why participants were or were not satisfied with their experience.

Table 3: Participant Satisfaction

Satisfaction Response Options	Number (n)	Proportion (%)
Extremely	3,890	15
Very	7,779	30
Neither satisfied or dissatisfied	1,297	5
Somewhat	5,186	20
Not at all	6,483	25
No response/missing	1,297	5
Total	**25,932**	**100**

Only 45% of program participants were very or extremely satisfied with the services offered.

50% of program participants were not at all, somewhat or neither satisfied nor dissatisfied with the services offered.

Low number of non-responses does not present a concern.

Outcome Analysis

Analyzing and interpreting program outcomes is a core part of an evaluation effort. Outcomes include short-, medium- and long-term changes the program produces in its participants in terms of functioning, behavior, symptoms, attitudes, knowledge, and other constructs related to program objectives. Below are common outcome evaluation questions that can guide program improvements.

Common Outcome Questions

Overall Effectiveness: ***Did the program achieve its intended outcomes?*** The answer to this question is critical in guiding program improvements, as well as in maintaining results over time. Specific outcomes are generally defined by program objectives aligned with its broader mission and goals. As such, it is critical that objectives are SMART (specific, measurable, achievable, relevant and time-bound) to set a clear reference point for the outcomes a program is expected to produce.

Group differences: ***Did the outcomes vary by subpopulation or intervention group?*** If so, this may mean the program's practices should be modified to better meet the needs of certain groups or that certain components within a program should be emphasized over others to meet the needs of different groups (e.g., groups exposed to different levels of stressors during deployments).

The example below demonstrates outcome analysis focused on two outcome domains of interest: resilience and learning. These outcome domains each have specific outcome measures. Analyses should be tailored to match the type of data derived from each of these measures.

Example: Outcome Evaluation

Program Sierra has a stated objective to improve resilience from baseline through a six-month follow-up. Measured outcomes revealed that average resilience ratings increased from 15 to 30 between baseline and post-program assessment but then declined to 20 by the six-month follow-up. As shown in the graph below, the error bars around the mean for

the baseline score do not overlap with the error bars around the mean of the post-program score. This supports the conclusion that increased resilience was indeed evident from baseline to post-program assessment. However, the benefit was not maintained over time. The fact that the error bars around the averages at baseline and those at six-month follow-up do overlap suggests that the scores are not significantly different.

This information may be very useful in terms of informing program improvements, which might include using additional training sessions to ensure sustained improvement or perhaps linking select participants to additional services.

Objective: Program participants will demonstrate improved resilience from baseline to post-program and will be maintained at six-month follow-up

Measured Outcome: On average, resilience ratings increased from 15 to 30, but then declined to 20

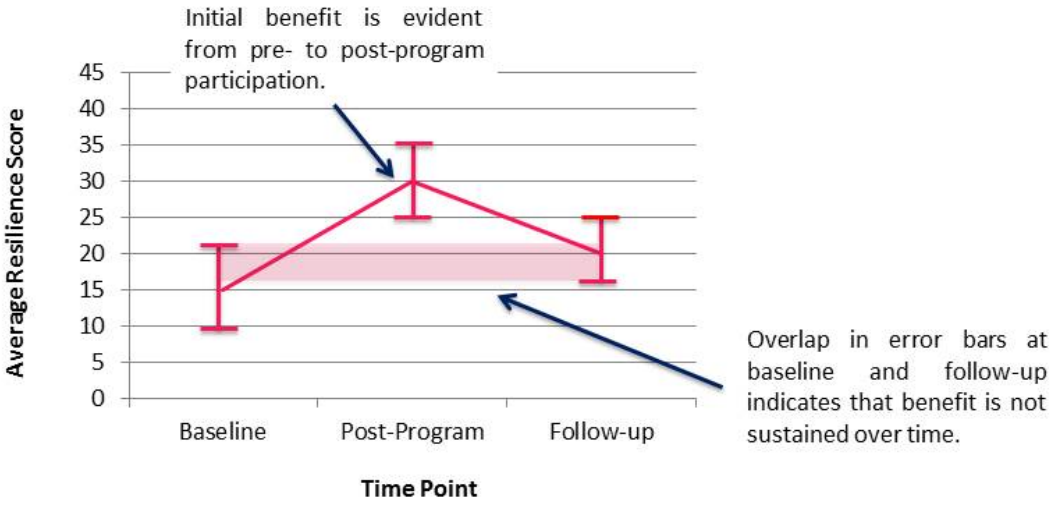


Figure 3: Effectiveness – Resilience Score Outcomes

Now consider a different type of outcome evaluation for Program Sierra focused on determining whether outcomes vary across sub-populations or intervention groups. For example, Program Sierra wants to examine how well one of their training initiatives is increasing learning among personnel across three different service branches. As seen in Figure 4, outcomes are similar between service branch A and service branch B, but the learning score is lower for service branch C. This informs a program manager about the need for modifications or improvements directed toward improving learning scores for service branch C. It could be that service branch C finds the subject matter irrelevant or that they cannot fully take advantage of the program because they are unable to attend consistently.

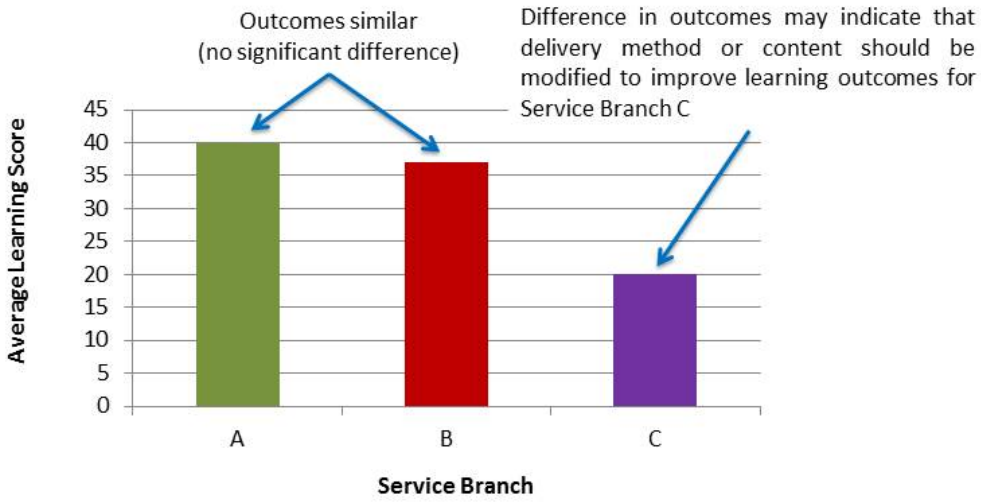


Figure 4: Difference by Groups – Learning Score Outcomes

As Program Sierra progresses through the outcome analysis process and is able to interpret and refine their findings, it will be important to present them to stakeholders in meaningful ways. Stakeholders want to know whether intended outcomes were achieved and how program managers know they were achieved. It is no longer enough to say, "We think we are benefiting participants." Rather, stakeholders want to hear, "The program is producing benefits, and we know we are achieving those benefits based on evidence from our ongoing program evaluation processes." Likewise, stakeholders want to know about multiple outcomes, including target outcomes and those that are the focus of their respective interests, such as maintaining force readiness or the ability of service members to perform their job functions. Measurement and analysis provide opportunities to highlight program strengths and to develop targeted program improvements based on actual program data.

Finally, stakeholders need details about how program improvements are to be carried out. Pointing to weaknesses or failings of a program is not particularly useful. However, by focusing the evaluation process on identifying opportunities for improvement supported by objective data, it is possible to define a clear path forward to better serve program participants and milestones for success.

Software for Quantitative Analysis

A variety of software packages exist to facilitate quantitative data analysis, ranging from traditional spreadsheet (e.g., Microsoft® Excel) and enhanced database processing applications (e.g., Microsoft Access® with Total Access Statistics) to more sophisticated free (e.g., R Programming) and commercial (e.g., IBM SPSS Statistics®, SAS®, or Stata®) software packages. Spreadsheets or enhanced database applications are generally appropriate for basic quantitative analyses, such as sorting, filtering, elementary descriptive statistics and creating visual representations and graphs. More advanced software packages include additional functionality that can provide inferential statistics capabilities such as the significance testing, model testing for theories and the ability to create visual charts that depict results. The choice of the computer software depends upon the requirements of the data plan and the complexity of the data analyses. Software packages may intend to make quantitative analyses easier to

perform, but keep in mind that more sophisticated packages will require more training and technical support.

Ensuring Quality in Results

As discussed in other PEG modules, validity and reliability are the two major dimensions of accuracy and precision used to support the quality of results from quantitative statistical methods and to reflect the multiple ways of establishing truth (Golafshani, 2003). Careful consideration for validity and reliability should be maintained throughout the entire evaluation process, including during quantitative analyses. Validity refers to the extent to which inferences, conclusions, and decisions made on the basis of collected data are appropriate and meaningful. Furthermore, validity defines the strength of the final results and whether they can be regarded as accurately describing the real world. For example, a program designed to improve depression symptoms should focus on quantitative data and outcome measures related to improvement of specific symptoms rather than general improvements in a participant's sense of wellbeing. There are three basic concepts used to determine the validity of collected data (Mason & Bramble, 1989): content validity, construct validity and criterion-related validity. These concepts and the means used to evaluate them are described in Appendix B alongside relevant concepts related to reliability.

Reliability refers to consistency of measurement (Bollen, 1989) or stability of measurement over a variety of conditions in which basically the same results should be obtained (Newcomer, Hatry, & Wholey, 2010). In simple terms, reliability describes the repeatability and consistency of assessments and measurements. For example, two different technicians conducting the same assessment/ procedure using the same tools/instruments with the same participant should yield the same or very similar results with the assumption that variables were maintained in a controlled environment. In this instance, the research instrument is considered to be reliable (Joppe, 2000). There are three basic methods used for estimating or verifying the reliability of collected data: test-retest, alternate or parallel form, and inter-rater/observer reliability.

Conclusion

At the conclusion of this module, Analyzing Quantitative Data, program evaluators should know the difference between descriptive and inferential statistical methods, and how to use them appropriately in describing, summarizing and comparing program information. Program evaluators should be able to initiate the analysis of program processes to show how their program actually operates, and use those results to connect processes to outcomes. This will help program evaluators gain a better understanding of how the program is impacting its participants and ultimately provide better understanding of what needs to be improved or changed in the program. Program evaluators should also seek to maintain high accuracy in their findings by attending to key dimensions such as validity and reliability.

Key Takeaways

- Quantitative methods can help evaluators explain the implications of observed data for larger populations and can also help address common process and outcome questions of interest
- Programs can use data analysis strategies to provide evidence of a program's effectiveness
- Data analysis can also be used to establish the degree to which a program's processes (inputs, activities, and outputs) contribute to its outcomes
- Analyses of program processes, outcomes and costs can help guide program improvements and support ongoing efforts toward maintaining desired results over time

References

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons, Inc.

Brent, D. (March 31, 2016). Data Storytelling: The Essential Data Science Skill Everyone Needs. *Forbes.* Retrieved from: http://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/#3f96436cf0c8

Brown, S. (2016). *Stats Without Tears – 10. Hypothesis Tests.* Retrieved from: http://brownmath.com/swt/chap10.htm

Coalition for Evidence-Based Policy. (January, 2014). Which Comparison-Group ("Quasi-Experimental") Study Designs Are Most Likely to Produce Valid Estimates of a Program's Impact?: A Brief Overview and Sample Review Form. Retrieved from: http://coalition4evidence.org/wp-content/uploads/2014/01/Validity-of-comparison-group-designs-updated-January-2014.pdf

Ellis, P.D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge, UK: Cambridge University Press.

Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*, *8*, 597-606. Retrieved from: http://nsuworks.nova.edu/tqr/vol8/iss4/6

Joppe, M. (2000). The Research Process, 1. Retrieved from: http://www.nova.edu/ssss/QR/QR8-4/golafshani.pdf

Mason, E.J., & Bramble, W.J. (1989). *Understanding and conducting research: Applications in education and the behavioral sciences.* New York, NY: McGraw-Hill.

Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (2010) Planning and designing useful evaluations. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.) *Handbook of practical program evaluation* (3rd ed., pp. 1-29). San Francisco: Jossey-Bass.

Selected Resources for Additional Study

Administration for Children and Families, Office of Planning, Research and Evaluation (2010). *The program manager's guide to evaluation* (2nd ed.).Retrieved from U.S. Department of Health and Human Services website: http://www.acf.hhs.gov/programs/opre/resource/the-program-managers-guide-to-evaluation-second-edition

Babbie, E. (2010).*The Practice of Social Research* (12th ed.). Belmont, California: Wadsworth, Cengage Learning.

Biddix, P. (n.d.). Research Rundowns. Quantitative Methods: Instrument, Validity, Reliability. Valdosta, GA: Dewar College of Education, Valdosta State University. Retrieved from: https://researchrundowns.com/quantitative-methods/instrument-validity-reliability

Bingham, R., & Felbinger, C. (2002*). Evaluation in Practice: A Methodological Approach* (2nd ed.). London, UK: Chatham House Publishers.

Carmines, E., & Zeller, R. (1986). Reliability and Validity Assessment. *Quantitative Applications in the Social Sciences Series*, 17(10). Albany, NY: Sage University Papers.

Centers for Disease Control and Prevention (2011a). *Introduction to program evaluation for public health programs: A self-study guide.* Retrieved from: http://www.cdc.gov/eval/guide/

Centers for Disease Control and Prevention (2011b). *Developing an effective evaluation plan*. Retrieved from: http://www.cdc.gov/obesity/downloads/CDC-Evaluation-Workbook-508.pdf

Creswell, J.W. (2013). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). Boston, MA: Pearson Education, Inc.

DCoE Program Evaluation Guide and Program Evaluation Webinar Series: http://www.dcoe.mil/About_DCoE/Program_Evaluation/Resources_and_Training.aspx

Defense and Veterans Brain Injury Center: http://dvbic.dcoe.mil/diagnosis-assessment?audience[0]=3

Deployment Health Clinical Center: http://www.pdhealth.mil/clinicians/assessment_tools.asp

Drost, E. (2015). Research and Perspectives. *Validity and Reliability in Social Science Research Education*, 38(1). Los Angeles, CA: California State University. Retrieved from: https://www.researchgate.net/publication/261473819_Validity_and_Reliability_in_Social Science_Research

Gusukuma, I. V. (2012). *Basic Data Analysis Guidelines for Research Students*. Belton, TX: University of Mary Hardin-Baylor.

Hair, Jr., J. F., Money, A. H., Samouel, P., & Page, M. (2007). *Research Methods for Business*. England: John Wiley & Sons Ltd.

Key, J. (1997). Research Design in Occupational Education. Module R10: Reliability and Validity. Oklahoma State University. Retrieved from: http://www.okstate.edu/ag/agedcm4h/academic/aged5980a/5980/newpage18.htm

Minnesota Department of Health: http://www.health.state.mn.us/divs/opi/qi/toolbox

National Center for Telehealth and Technology: http://t2health.dcoe.mil/

National Network of Libraries of Medicine: http://nnlm.gov/evaluation/guides.html

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Peat, J. (2002). *Health Services, Research: A Handbook of Quantitative Methods*. London, UK: Sage.

The Community Tool Box, University of Kansas: http://ctb.ku.edu/en

Torres-Reyna, O. Data Preparation & Descriptive Statistics (PowerPoint slides). Princeton, NJ: Princeton University, Data & Statistical Services. Retrieved from: http://www.princeton.edu/~otorres/DataPrep101.pdf

Trochim, W. (2006). Research Methods Knowledge Database: Construct Validity. Web Center for Social Research Methods. Retrieved from: http://www.socialresearchmethods.net/kb/constval.php

U.S. Army Public Health Command, Behavioral and Social Health Outcomes Program (BSHOP): http://phc.amedd.army.mil/topics/healthsurv/bhe/Pages/BehavioralandSocialHealthOutcomesProgram%28BSHOP%29Services.aspx

University of Kentucky – Extension: http://www2.ca.uky.edu/AgPSD/Focus.pdf

Appendix A. Fidelity – Process Analysis Matrix

This matrix is an example of a process analysis tool an evaluation team can use as a guide in their efforts to determine whether their program was implemented with fidelity.

Metrics	Implementation Status	Current Status	Degree of Change
Coverage	What percent of the target population was covered by Program Sierra at implementation?	What percent of the target population is currently being covered by Program Sierra?	Has coverage increased, decreased or remained unchanged? Why?
Content	What activities were conducted at implementation?	What activities are currently being conducted?	Are the same activities being conducted that were implemented?
Frequency	How frequently were activities being conducted at implementation?	How frequently are activities currently being conducted?	Has the frequency of each activity remained the same or changed over time?
Duration	What was the duration of each activity at program implementation?	What is the duration of each activity currently?	Has the duration of each activity remained the same or changed over time?

Appendix B. Dimensions of Validity and Reliability

This table provides some examples of quantitative program evaluation practices that support each of the methods listed for validity and reliability. These practices can be used to enhance the quality of collected data and the credibility of results and findings.

Method	Definition	Evaluation Practice
Content Validity	Extent to which data represents the domain or universe of the trait or property being measured	Identify the overall pool of content to be represented. Then choose items randomly from the pool that will accurately represent the information in all areas.
Construct Validity	Extent to which theories and observations can be matched and translated into actual programs or measures	Clearly define program mission using SMART objectives as it will form the foundation and focus of assessment tools and measures.
Criterion-Related Validity	Extent to which data can be used to predict future or current performance (i.e., the correlation of results with another criterion of interest)	Compare program evaluation data and performance results with similar programs to determine need for refinement of assessment tools and measures.
Test-Retest Reliability	Extent to which data is consistent and stable when compared or evaluated over time	Conduct a series of pilot tests for assessment(s) to determine whether data results are consistent and whether varying results highlight inconsistencies.
Parallel Form Reliability	Extent to which two assessments constructed the same way from the same content provide consistent data results	Use alternate assessments that address variations of understanding and interpretation among the target population (e.g., language barriers, cultural origins).
Inter-Rater/ Observer Reliability	Extent to which different raters/observers give consistent answers or estimates	Conduct training for raters/observers that clearly defines rating criteria and provides standardized rating forms/tools.