



Planning for Program Evaluation: Develop Evaluation and Measurement Strategies

Presented on Feb. 17, 2015

Episode 3 in the FY2015 Program Evaluation and Improvement Training Series

Presenters:

CAPT Armen Thoumaian, Ph.D., USPHS
Health Science Officer
Office of Shared Services Support
DCoE

Aaron Sawyer, Ph.D.
Research Scientist
Contract support for DCoE

Richard Best, Ph.D.
Research Scientist
Contract support for DCoE

Moderator:

Carmina Aguirre, M.A.
Research Scientist
Contract support for DCoE

[Video Introduction]

CAPT Thoumaian: Hello. My name is Captain Armen Thoumaian of the Defense Centers of Excellence for Psychological Health and Traumatic Brain Injury or DCoE. Thank you for joining us for this episode of the DCoE Program Evaluation and Improvement webinar training series.

DCoE's Mission is to improve the lives of our nation's service members, families and veterans by advancing excellence in psychological health and traumatic brain injury prevention and care.

DCoE accomplishes that mission in coordination with its three Centers: Defense and Veterans Brain Injury Center, Deployment Health Clinical Center and National Center for Telehealth and Technology. Together, we produce a variety of trainings on subjects ranging from program evaluation to clinical care and prevention practices.

This training series is designed for program administrators and service leadership who are

involved with or who plan to conduct program evaluation activities within the Defense Department's psychological health and traumatic brain injury programs. Our objective is to enhance the capability of these personnel to actively engage in program evaluation activities and, ultimately, make program evaluation an inherent component of everyday program operations.

By supporting enhanced program evaluation capabilities across the Defense Department, this series contributes to DCoE's larger mission to improve the quality and effectiveness of the psychological health and traumatic brain injury prevention and care programs that serve our military members, their families and veterans.

On behalf of DCoE, thank you for participating in this training series.

[Slide 1]

Ms. Aguirre: Hello. My name is Carmina Aguirre. I provide contract support to the Defense Centers of Excellence for Psychological Health and Traumatic Brain Injury or DCoE. I will be your moderator for this presentation, the third episode in the 2015 DCoE Program Evaluation and Improvement webinar training series. The webinar is hosted using the Adobe Connect platform, and the technical features are being handled by DCoE's webinar support team in Washington, D.C.

Today's topic is "Planning for Program Evaluation: Develop Evaluation and Measurement Strategies." Before we begin, let's review some details.

[Slide 2]

This presentation has been pre-recorded; however, there will be a live Question-and-Answer session at the end of the presentation.

Throughout the webinar, we encourage you to submit technical or content-related questions using the Question pod on your screen. Your questions will remain anonymous, and our presenters will respond to as many questions as possible during the Q-and-A.

At the bottom of the screen is the Chat pod. Please feel free to identify yourselves to other attendees and to communicate with one another. Time is allotted at the end of the presentation to use the Chat pod for networking.

All audio is provided through the Adobe Connect platform; there is no separate audio dial-in line. Please note there may be delays at times as the connection catches up with the audio. Depending on your network security settings, there may also be some noticeable buffering delays.

Closed captioning is provided for today's event, and a transcript will be made available at a later date.

[Slide 3]

Webinar materials for this series are available in the Files pod at the bottom of the screen during the webinar. They are also posted in the Program Evaluation section of the DCoE website. Modules from the newly revised DCoE Program Evaluation Guide will be posted throughout the

2015 webinar series.

For information about other DCoE webinars and trainings, visit the Training section of the DCoE website by following the link on slide 3.

[Slide 4]

We are pleased to offer continuing education credit for the 2015 Program Evaluation and Improvement webinar series. Instructions for obtaining continuing education were made available during the registration process. Eligibility criteria for continuing education credit are presented on slide 4.

[Slide 5 through 8]

If you preregistered for the webinar and want to obtain CE certificates or a certificate of attendance, you must complete the online CE post-test and the evaluation. After the webinar, please visit continuingeducation.dcri.duke.edu to complete the online CE post-test and evaluation and download your CE certificate or certificate of attendance. The Duke Medicine website online CE post-test and evaluation will be open through February 24th, 2015, until 11:59 p.m. Eastern Time. Additional details regards continuing education can be found on slides 6, 7 and 8.

[Slide 9]

This webinar was introduced by Captain Armen Thoumaian. Captain Thoumaian is the Deputy Chief of Integration for the Office of Shared Services Support at DCoE. He is a Scientist Director in the Commissioned Corps of the U.S. Public Health Service with more than 30 years of experience in health and mental health program design and evaluation. In January 2012, Captain Thoumaian joined DCoE to help design and implement program evaluation and improvement efforts in the Defense Department. He holds a B.A. in psychology and sociology, an M.A. in general experimental psychology, and a Ph.D. in social welfare and social work. Captain Thoumaian has also completed a National Institute of Mental Health fellowship in Community Mental Health.

[Slide 10]

Our first presenter is Dr. Aaron Sawyer. Dr. Sawyer is a research scientist who provides contract support to DCoE. He is a clinical psychologist with extensive expertise in intervention outcome research and program evaluation. He has delivered child, family, and adult interventions for more than a decade, including specialization in trauma and experience working with military families. Dr. Sawyer holds a master's degree in experimental psychology and a doctorate in clinical psychology. He completed postdoctoral training at The Kennedy Krieger Institute of Johns Hopkins University and is a licensed psychologist.

Our next presenter is Dr. Richard Best. Dr. Best is an industrial and organizational psychologist with 14 years of experience conducting health services research in both the Veterans Health Administration and the Defense Department's Military Health System. He has extensive experience in research design, qualitative and quantitative data collection and analysis, and collaborating with clinical experts to translate research results into actionable recommendations. Dr. Best holds a master of science and a Ph.D. in industrial-organizational psychology and is certified in Prosci's change management process.

[Slide 11]

I am Carmina Aguirre, your moderator for today. I am also a research scientist who provides contract support to DCoE. I have over 14 years of experience within the Defense Department. My background includes executive leadership, psychological health, sexual assault prevention and response and public affairs. In addition to supporting DCoE, I serve as Chief of Public Affairs in the Florida Air National Guard. I hold a B.A. in psychology and an M.A. in human services with a specialization in executive leadership.

[Slide 12]

This training presentation will provide guidance on selecting evaluation designs and questions to meet a program's evaluation goals. In addition, it will describe important considerations for selecting measurement strategies and metrics.

At the conclusion of this webinar, participants will be able to:

- Choose an appropriate evaluation design and develop evaluation questions
- Explain how metrics and measurement strategies are used in program evaluation and improvement efforts
- Demonstrate knowledge of important considerations for selecting or developing measures
- Select and implement strategies to address common measurement challenges

[Slide 13]

As seen on slide 13, Captain Thoumaian will begin with a discussion of designing a program evaluation. Dr. Sawyer will then provide an overview of key concepts related to measurement as well as a discussion of how to choose general areas to be measured. Dr. Best will then discuss measure selection and development, followed by ways to overcome common challenges. We will conclude with a summary by Captain Thoumaian. Then, I will provide a list of references and resources, followed by a question-and-answer session with our presenters and you will have an opportunity to provide anonymous feedback.

[Slide 14]

CAPT Thoumaian: Thank you, Ms. Aguirre. In this section, I will describe important considerations for how to design program evaluations, including the creation of specific evaluation questions. Careful selection of an appropriate design and questions at the outset of an evaluation effort will help to ensure its success.

[Slide 15]

Peter Drucker, an influential management consultant, educator and author, said, "The most serious mistakes are not being made as a result of wrong answers. The truly dangerous thing is asking the wrong question."

This quote reflects the importance of good planning during the early stages of a program evaluation effort. That is, you have to choose the right approach and the right questions in order to get to the right answers. Because programs are complex and have numerous moving parts, there are endless aspects that could be evaluated. However, through careful consideration of what may be learned through an evaluation, it is possible to select the right starting point and proceed forward.

[Slide 16]

Program evaluations often have similar purposes and designs to ensure comparability between programs and within a single program over time. However, for evaluations that are internal to a program, it is often appropriate to tailor an evaluation design to meet the needs of a specific program evaluation. It is important that the evaluation design be established early in the process to ensure that all that appropriate resources are acquired to complete necessary tasks.

A program's specific approach to evaluation should be determined by three primary factors, which will be described in more detail on the following slides:

1. First, program evaluators must consider the goals of an evaluation, which are determined by key stakeholders such as leadership members who oversee the program, staff who carry out a program's activities and representatives of the groups who actually participate, such as service members and their families.
2. Second, the nature and intent of the program will guide evaluation efforts. These are defined in program objectives and a logic model that outlines how the program is organized.
3. Third, a program's level of maturity, also known as its stage of development, must be taken into account, because programs with varying levels of maturity often have different organizational capacities and capabilities.

[Slide 17]

Evaluation goals should be determined prior to executing any evaluation activities. Evaluation goals are established early in the evaluation process by communicating with stakeholders to determine their specific needs and interests in the evaluation effort.

"Stakeholders are people or organizations that are invested in the program, are interested in the results of the evaluation and/or have a stake in what will be done with the results of the evaluation." This includes not only decision-makers and funding agencies but also those who work within the program and those who are served by it.

For accountability purposes, stakeholders often want to know to what extent a program is achieving its intended outcomes and how efficiently resources are being used. In addition, they may be interested in big-picture matters, such as how the program fits into the overall system of prevention and care or how the program helps the military as a whole to achieve its mission. However, stakeholder interests vary widely, as described in Module 3 of the Program Evaluation Guide.

Regular communication with stakeholders throughout the evaluation process helps sustain support for the evaluation effort. Consequently, we strongly encourage continuous engagement

with stakeholders as part of any program evaluation and improvement effort.

[Slide 18]

The second factor that determines evaluation design is the nature and intent of a program. These were also discussed in detail in the previous episode in this series and in Module 2 of the Program Evaluation Guide.

A program's mission, goals and objectives describe in increasingly specific detail what a program is intended to achieve. Evaluations are designed to assess whether a program is achieving its objectives and how it does so, or if not, what can be improved upon to enhance quality and effectiveness.

A program logic model demonstrates how a program's resources, or inputs, are used to conduct the program's core activities in order to produce outputs and outcomes. The logic model is informative in guiding evaluation design, because an evaluation generally focuses on specific parts of a program, such as achievement of outcomes or whether activities are conducted as planned.

[Slide 19]

The third factor to consider in designing a program evaluation is program maturity, for which the CDC describes three stages:

1. One, programs in the planning stage are just getting started. Their activities are as yet untested, and the focus of evaluations at this stage is primarily on refining program plans.
2. Two, programs in the implementation stage are operational but still adapting, much as teenagers are fairly independent but perhaps not quite ready to be on their own. The evaluation focus during this stage is typically on refining and improving operations. At this stage, a program can be expected to produce information about its use of resources as well as its activities and outputs.
3. Three, programs in the outcomes stage should be able to produce desired results and also be able to show that those results are indeed being achieved. Consequently, the focus of evaluations in this stage is often on measuring those results and also determining whether any unintended results have occurred.

The divisions between these stages are not clear-cut. Rather, different aspects of a program may be in different stages, such as when implementing a new practice within a mature program. Likewise, program personnel may wish to focus their evaluation effort on multiple areas at the same time.

[Slide 20]

Based on evaluation goals, the nature and intent of a program, and program maturity, you may choose from three broad types of evaluations designs:

- Formative evaluations are used during program planning to assess the need for a program to be developed, and whether a program has important structures in place like

funding, staffing and reporting capabilities.

- Process evaluations are used during program implementation to determine how well the program is operating, and to guide refinements to program activities.
- Summative evaluations are used with programs that have been established and operating for some time to determine whether they are effective.

In the slides that follow, we describe in more detail subtypes, or areas of focus within each of type of evaluation design, as well as specific evaluation questions that may be addressed.

[Slide 21]

The focus of formative evaluation designs may include the needs of a population that call for a program to be developed, or the resources required to set up and operate a program over time. Alternatively, an evaluability assessment may be conducted to determine whether a program is capable of producing enough information to actively participate in a more thorough evaluation.

In the box at the bottom of slide 21, we provide sample questions from Module 3 of DCoE's Program Evaluation Guide. We encourage you to read through these questions in more detail and consider what kinds of questions you might ask in an evaluation of your program, taking into account evaluation goals, the nature and intent of the program and program maturity. Note that an evaluation effort may address multiple questions that cut across different evaluation designs.

As an example of a formative evaluation question, consider an initial evaluation question, "Can the program be implemented?" An evaluation following from this question might gather and analyze data about whether the program has the necessary resources to operate, such as funding, space and personnel with appropriate training. Similarly, examine the middle question on the right, "Does the program have the structures in place to be evaluated?" This refers to an evaluability assessment. This is important in determining a program's capacity to actually complete an evaluation, and to generate the information needed to answer advanced evaluation questions for a process or summative evaluation.

[Slide 22]

For process evaluation designs, a common area of interest is fidelity, or the degree to which a program operates according to a plan, such as a set of best practices. Likewise, a program may wish to examine whether it provides services, or coverage, to its target population. The results of process evaluation designs can be used to refine program operations.

Consider the first process evaluation question listed below: "How similar are participants to the target population for which the program was designed?" An evaluation following from this question would compare the measured characteristics of participants, such as their demographics and presenting conditions, to the target population identified. Likewise, the last question on the right reads, "Are participants being followed during and upon conclusion of program services?" To address this question, evaluators must assess the capacity of a program to track participant outcomes and perhaps to track service utilization after program completion.

[Slide 23]

Summative evaluation designs focus on the overall results of the program, that is, whether the program accomplishes its mission, goals and objectives. Outcome evaluations are a type of summative evaluation focusing specifically on whether a program achieves desired changes among program participants. For example, reducing symptoms or learning new skills. Other types of summative evaluations may focus on whether outcomes can be attributed to program activities, or whether a program's benefits are worth the costs of operating. Summative evaluations are generally applied to more mature programs that have been in existence long enough to stabilize their operations and collect outcome information.

In the box below, some summative evaluation questions are obvious, like the top one, "To what extent did the program achieve the desired outcomes?" However, there are also more subtle questions, such as, "Did the program impact vary across groups?" The latter question requires that programs be able to tie outcome information to different groups, such as men versus women or active-duty versus the reserve components. In addition, another important question, second from the bottom on the right, reads, "Does the benefit of the program warrant its costs?" This type of question is of great interest to funding agencies and would require detailed information about costs relative to outcomes.

[Slide 24]

To help you apply the information just discussed, we revisit Program Sierra¹, a hypothetical reintegration program discussed in the previous episode and in Module 2 of the revised Program Evaluation Guide.

To review, Program Sierra seeks to ensure that service members who are wounded, ill or injured successfully reintegrate into civilian life or return to active-duty in the military. By performing its mission effectively, Program Sierra personnel hope to enhance force readiness and improve the quality and efficiency of services across the Defense Department.

For easy reference, we have also provided Program Sierra's mission, goals, SMART objectives and logic model in the extra slides at the end of this presentation.

[Slide 25]

Earlier, I stated that three factors drive decisions about program evaluation design: evaluation goals, a program's nature and intent, and program maturity.

Program Sierra's leadership and stakeholders discussed their evaluation goals and determined that they are reasonably confident that the program is reaching its target population. However, they wanted to know, first, how well the program is being implemented, and second, whether program activities are actually leading to the expected outcomes.

Program Sierra's nature and intent are described in detail in its SMART objectives and logic model, provided on slides 73 through 77 [see slides 69 through 74]. A key point to mention here is that the nature and intent are clearly specified in objectives and the logic model, which provide the foundation for focused evaluation questions and measurement.

¹ Program Sierra was formerly known as Program Echo.

In terms of program maturity, Program Sierra is in the implementation stage – it has already been operating for a few years. However, because evaluation was built into the program from the outset, it has some of the outcome information needed to address the second evaluation goal, whether program activities lead to expected outcomes.

Given this information, take some time to consider what type of evaluation design is most appropriate.

[Slide 26]

The most appropriate evaluation design, overall, is a process evaluation specifically focused on Program Sierra's direct services to participants. However, since short-term outcome data are available, it is possible to incorporate some aspects of a summative evaluation design.

Relevant evaluation questions help to focus the evaluation effort further. Specifically, Program Sierra's evaluation team, led by its program manager, will examine whether the program was implemented with fidelity, and whether it achieved short-term outcomes. Based on the answers to these questions, the team can determine what should be improved within the program to enhance quality and effectiveness.

We encourage you to go through this same exercise with your own program, as application of these concepts is the best way to learn about evaluation.

[Slide 27]

To summarize this section, choose as a primary focus of your evaluation effort one of the three types of evaluation designs: formative, process or summative. Then, develop more focused evaluation questions. Rather than limiting the scope of an evaluation effort to a single area, you may wish to include aspects of multiple designs when feasible. Choose an evaluation strategy that matches evaluation goals, rather than goals to fit the strategy.

By being diligent about evaluation planning, you will be better equipped to accomplish evaluation goals. In addition, proper planning will help to ensure that goals are accomplished on time and with minimal impact on program operations.

Now, Dr. Sawyer will describe key concepts underlying measurement in program evaluation.

[Slide 28]

Dr. Sawyer: Thank you, Captain Thoumaian. In this section, I will provide an overview of important concepts to keep in mind as you design and carry out an evaluation.

[Slide 29]

Nearly a hundred years ago, pioneering behavioral psychologists E. L. Thorndike said, "Whatever exists at all, exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality."

I studied measurement and evaluation under his grandson Robert M. Thorndike, a towering man with a booming voice and a deep love of measurement. Like his grandfather, his father Robert L. Thorndike was a psychologist; and his daughter is a psychologist too. While many of

us likely grew up with jobs such as cutting grass and babysitting, Robert M. Thorndike grew up spending his summers on a boat in the Pacific Northwest with his father creating intelligence test questions.

Around the time I studied with him, Robert M. Thorndike was completing a large landscaping project at his home. He suspected that the stone yard that provided his landscaping rocks had 'shorted' him on his two-ton order. So, as a lover of measurement, he began weighing stones. Over the course of several days, he discovered his order was indeed hundreds of pounds short. He confronted the stone yard manager with his findings, and you can only imagine the manager's face before he began loading the missing stones in a truck along with a few hundred pounds extra for good measure.

The point to this story is that unless you accurately measure something, you don't really know for sure if your results are as expected. Our stakeholders love measurement almost as much as the Thorndikes, and they expect that if you tell them your program has achieved certain objectives, you can back up those statements with accurate data.

[Slide 30]

Measurement is the process of collecting information, or data, about some area of interest. Often, people think of data in terms of numbers, or quantitative data. However, data can be any type of information, including text, voice or video recordings, figures or drawings, and other non-numeric information that may be generated as part of an evaluation effort.

A metric is the standard by which something is measured. For example, we use the imperial measurement system in the U.S. to measure distance, whereas in most of the world, the metric system forms the standard. Similarly, temperature may be measured in Fahrenheit, Celsius or Kelvin. These are measurement standards.

Finally, a measure is a specific tool used to collect data. For example, a ruler might be used to assess length, just as a specific questionnaire could be used to assess posttraumatic stress disorder symptoms or quality of life. In practice, the terms metric and measure are often interchangeable.

[Slide 31]

In classical measurement theory, a given data point, X , is the combination of a true value, T , and some amount of measurement error, E . So, let's say I'm measuring risk for suicide by having service members complete questionnaires as they return from deployment. If I give a service member the questionnaire form, I will get a risk score. That score is related to the true amount of risk but also contains some error. Error, in this case, may be present because the service member had difficulty concentrating, because the questions were difficult to understand, or perhaps because the service member did not want to respond in a way that might delay the trip home or affect his or her career.

Even the best measurement will have some degree of error. The goal for those conducting evaluations is to minimize error so that measured data points are as close to the true value as possible.

[Slide 32]

Validity and reliability are closely related to the amount of error involved in conducting measurement. Validity is the degree to which a measure accurately represents the characteristic it is designed to measure. Reliability refers to whether the results of measurement are consistent across time and situations.

If I want to assess risk for a future heart attack, a valid measure might include a combination of ratings based on blood pressure, family history, cholesterol level, exercise, diet, weight and health habits, all of which have been shown to predict heart attacks. A measure containing those risk factors would allow me to produce a valid estimate of a person's risk for a heart attack.

On the other hand, I could measure height as a risk factor for heart attacks, and that would be pretty reliable since height doesn't change much throughout adulthood. Height would be reliable but not valid because there is no research to show that height is related to heart attack risk.

It is important that measurement strategies be both valid and reliable, and as we will discuss, using multiple measurement strategies and data sources will help to overcome the limitations of any one strategy or source.

[Slide 33]

Qualitative data are non-numeric forms of data such as text, audio and video recordings, or pictures. Qualitative data are widely used in program evaluation efforts. They are often referred to as "soft" data, while quantitative data are commonly referred to as "hard" data. However, qualitative data can provide equally valid measurements of characteristics or situations, even though they may be more context-specific and therefore less reliable.

Some of the most commonly-used qualitative data types are listed on slide 33.

- Interviews in a one-on-one situation may be conducted with program providers, program participants and others who may offer valuable insights into how a program actually functions. One-on-one interviews yield highly detailed information, since the interviewer can ask follow-up questions and gain an in-depth understanding of the difference a program makes in people's lives. Interviews vary in their degree of structure, or how scripted the interview process is.
- Focus groups use a leader to guide a structured discussion among 4 to 12 people. The leader should be someone with focus group experience. Generally speaking, the richest information is obtained when there is a free-flowing discussion among focus group participants. To that end, it is advisable that focus groups consist of individuals who are similar to one another in terms of key characteristics like status or rank.
- Open-ended comments are used quite often and include written responses on forms or surveys, such as the comment box on the Interactive Customer Evaluation card we ask you to complete at the end of this webinar. These comments include thoughts, opinions and suggestions for improvement.
- Observations are used to document an activity, process or conditions of a facility. Very often a checklist is employed so that observers record what they see in a consistent manner and spend more time observing than writing. Of course, you must be aware of privacy concerns when using this method and also note that the act of observation may

cause those being observed to act differently.

- You may be very familiar with After Action Reviews, or AARs, or other types of process reviews in which staff members who participated in an activity discuss its strengths and weaknesses. These are also known as “hotwashes.”
- Case studies are a very useful way to present data about how a program works—or where it does not work—from the participant’s point of view. Case studies gather data over time, starting from entry into a program through the program’s end and perhaps beyond. This type of data draws attention to the program in a detailed and compelling way.

Ways to collect, code and use qualitative data will be presented in a future episode in this training series.

[Slide 34]

On slide 34, there are four basic types of quantitative or numeric data. Nominal data refer to categories, such as gender, ethnicity and service branch. In a dataset, you might use a zero to represent men and a one for women, but the difference is a matter of type rather than women being one unit greater than men in some quality.

Ordinal data are rank order, such that there is an order to different numbers, but the difference between a rank of one and two isn’t necessarily the same size as the difference between a rank of two and three. Many forms of data are ordinal, such as opinion surveys and many behavioral rating scales.

Interval and ratio data do have equal units, so the difference between one and two is the same amount as the difference between two and three. The only distinction between interval and ratio data is that ratio data have a real zero point. So, reaction time is ratio, since it can never be negative, whereas time measured on a clock is interval, because we set the zero point arbitrarily at midnight.

These distinctions will become clearer as we discuss data collection and analysis in future episodes in this series.

[Slide 35]

Qualitative and quantitative methods should not be considered as competing ways to gather information for an evaluation effort. In fact, it is generally advisable to use multiple methods to monitor program quality and effectiveness.

Mixed methods refer to the combined use of both qualitative and quantitative evaluation strategies. Mixed methods offset the weaknesses of quantitative and qualitative approaches by drawing upon the strengths of each. They may also allow for the exploration of different types of evaluation questions, which may be better suited to one method or the other.

Mixed methods may be used simultaneously or sequentially, depending upon evaluation needs. For example, mixed methods might be used to assess the size of changes in traumatic brain injury symptoms or decreases in the frequency of problematic behaviors like substance abuse, while also looking at how participants understand their challenges and their interactions with

treatment providers.

Likewise, mixed methods can be used to answer multiple different questions in a broader evaluation effort. For example, focus groups could be conducted with program staff and participants to identify ways to improve the relevance of content for a web-based resiliency training program. Meanwhile, quantitative methods could be used to determine the degree of learning improvement from baseline to after the training content is delivered.

[Slide 36]

Before making choices about specific measures and metrics in an evaluation, you must determine, more generally what areas are of interest. That is, what do you want to measure as part of your evaluation effort? Later, we will discuss how to measure those areas of interest.

[Slide 37]

Objectives are the most specific type of statement about the intent or purpose of the program, and a program logic model lays out how the program is organized to achieve objectives. Together, these play a key role in narrowing down what should be measured as part of an evaluation effort. This is where the work of developing SMART objectives and logic models pays off.

If you have well-developed SMART objectives and a detailed logic model, choices about focal areas for measurement can be as simple as pointing to or circling elements within them and deciding what metrics or measures to apply. Evaluation questions will direct your focus.

[Slide 38]

On slide 38, you see an example of how program objectives may be used to identify areas for evaluation measurement. Objectives should describe how inputs will be used, what activities will be conducted, what outputs will be produced and/or what outcomes are expected to result. Program evaluations will focus on one or more of these areas and the connections between them.

In the example on slide 38, consider a web-based training on posttraumatic stress designed for a broad audience. The stated objective is that program staff will provide monthly web-based training to unit commanders and enlisted personnel, who will demonstrate increases in their knowledge of posttraumatic stress symptoms between pre- and post-training assessment.

So, if a goal of your evaluation is to determine whether this objective is being achieved, you may wish to measure the following:

1. First, measure whether the web-based trainings are in fact being delivered as planned.
2. Second, you could measure who is actually participating in the trainings. That is, are unit commanders and enlisted personnel attending, or is it more one group than the other?
3. Third, with respect to outcomes, do participants actually gain knowledge, based on comparisons of pre-training and post-training assessments?

Keep in mind, there are literally hundreds of possible evaluation questions that could be

addressed, and this is just one sample objective.

[Slide 39]

Slide 39 depicts the core components of a logic model discussed in the previous episode and in Module 2 of the Program Evaluation Guide. Note that assumptions and external factors, although important, are not included here because they are generally not a focus of program evaluations.

For a quick review, inputs are the resources a program requires to operate, activities are what program personnel do in service to the program's mission, outputs are the products and participation that result from activities, and outcomes are the changes that result among program participants.

An actual program logic model should be arranged to match important goals and objectives. In this basic depiction, however, you can see a few of the major areas that would be included within each component. In general, a program should be measuring every element that shows up in a logic model on an ongoing basis. For the purposes of a specific evaluation, however, the focus might be on a more select set of program elements, such as finances from the input component, clinical and outreach activities from the activities component, products and participation from the outputs component, and short- to medium-term outcomes from the outcomes component.

As a useful application exercise, take a few moments to write down some possible evaluation questions you would be interested in examining for your own program. Then, follow up by examining your program's objectives and logic model to determine what areas you would have to measure to get to the answers or results.

[Slide 40]

On the next several slides, we discuss more specific examples of metrics within each major logic model component. We will highlight just a few examples from each. These are often important metrics to track on an ongoing basis, and they may be assessed using quantitative or qualitative methods, or in some cases both.

Inputs may be tracked using forms, records or logs in order to examine areas such as staffing, materials and finances. To highlight one area, staffing metrics may include:

- staff demographics, such as gender, ethnicity and education
- staff turnover from year to year
- qualifications such as licensure, certifications or completion of required trainings

These metrics in turn can be used to answer a variety of questions, such as whether the program has the staffing resources it needs to carry out key activities.

[Slide 41]

Moving on to the next component of the logic model, be sure to track all major program delivery activities. Often this will mean documenting that activities occur and also any outputs of those

activities. Tracking activities is an important way to demonstrate accomplishments.

For example, program personnel should track their delivery of core services, whether that consists of webinars, classroom-based trainings, clinical treatments or some other format. Likewise, promotional or outreach activities may be tracked, such as handing out fliers or staffing a health fair booth. Other activities may facilitate relationships with other programs or the scientific community, such as conference attendance or publication of white papers or peer-reviewed articles.

[Slide 42]

Also track information about program participants and their attendance as well as products of program activities. For example, a web-based program might measure participation in terms of downloads or web hits, whereas a classroom-based training would measure physical participation. When possible, it is useful to track demographic characteristics of participants to determine whether or not the program is reaching its intended audience.

Products are the tangible outputs of program activities. For instance, a program may track the number of pamphlets or fliers distributed, the number of webcasts delivered or the number of service units provided.

[Slide 43]

Of course, tracking outcomes is very important to evaluation efforts. Outcomes include short-term metrics such as changes in awareness or knowledge, medium-term changes such as increased use of coping skills or memory function, and long-term changes such as improved unit readiness or changes in norms and quality of life.

[Slide 44]

Measurement strategies are most effective when they use multiple methods and multiple sources of information for each area of interest. Rather than being redundant, multiple approaches can provide complementary or differing viewpoints. In addition, multiple approaches help the evaluator to overcome some of the limitations of any single measurement strategy. Likewise, by conducting measurement across multiple areas of interest, an evaluator can better capture the full breadth of how a program works and how it affects participants. For this reason, we generally recommend two to three metrics for each area of interest. In the figure on slide 44, you see an example of how this approach applies to outcomes.

The figure shows three possible outcome areas that could be measured as part of a program evaluation effort for a resiliency training program. Each outcome area includes two metrics that vary in source of information. Resiliency, for example, is measured by providing a self-report questionnaire to program participants, and also by conducting focus groups with participants in which they discuss the benefits they received from the program and areas to improve its services.

Job functioning, another outcome area of interest, is measured by days of work missed per year gathered from administrative records, and also by brief interviews with unit commanders about participating service members' job performance.

Finally, family relationships are measured by questionnaires provided to participants and their

spouses or partners, as well as individual interviews with spouses or partners.

By measuring multiple areas using different strategies, evaluators are able to comprehensively examine evaluation questions from varying angles. This same concept applies beyond outcomes in many cases to include outputs, activities and inputs.

[Slide 45]

The end result of the planning phase of an evaluation effort is a comprehensive plan for data collection and storage. This data plan should include information about what will be collected, by whom and when, how data will be stored and analyzed, and how quality assurance will be carried out to ensure that data are accurate.

Details on data collection will be provided in the next episode in this series.

[Slide 46]

Slide 46 presents a data matrix template. A data matrix is basically a table representing key aspects of the data plan. You may wish to use this template to assist you in specifying the details of a data plan. Note that multiple data matrices may be included for different areas of interest.

[Slide 47]

Slide 47 presents a data matrix example focused on short-term outcomes for Program Sierra, the hypothetical reintegration program discussed earlier and in the previous episode.

The Outcome Metric 1 column shows that service providers collect questionnaires from participants on their attitudes toward reintegration and ability to manage challenges. This information is to be collected before and after the intervention is delivered and at 3-month follow-up. A program manager is given responsibility for analyzing data for all metrics, and this particular metric will be used for both intervention planning and outcome tracking.

The remaining two columns show key details about other data sources including focus groups used to assess program benefits and areas for program improvement, as well as provider ratings of gains exhibited by participants. Similar examples will be presented as part of forthcoming Program Evaluation Guide modules.

And now Dr. Best will continue with a discussion of measure selection and development, which will inform the details of a data plan.

[Slide 48]

Dr. Best: Thank you, Dr. Sawyer. This portion of the webinar will provide guidance on selecting or developing measures, which is important when considering how to measure outputs and outcomes.

As discussed above, measurement strategies are most effective when they use multiple methods and informants or data sources. This minimizes the amount of error or bias in any single method or informant.

It is difficult to truly achieve the quality of measurement desired by researchers and psychometricians, or the people who study measurement, but it is important for program administrators and staff to aim high to get the most accurate measurements possible.

[Slide 49]

To select or to develop a measure, that is the question...

In large part, this question is about which metric or metrics are the most valid and reliable for a program's evaluation goals, the services it provides and the population it serves. Generally speaking it is best to select existing measures with proven reliability and validity. There are a number of resources that can guide the selection of measures for specific programs, including those listed in the References and Resources section of this webinar. Examples include the Agency for Healthcare Research and Quality, the National Quality Forum, the National Center for PTSD and DCoE's Centers – DHCC, DVBIC and T2.

New or custom measures are most appropriate when assessing learning outcomes relevant to a specific program process, such as a skills training class or participation in a webinar. However, there are a number of best practices and caveats to consider when developing a new measure, as will be discussed shortly.

It is often useful to consult with experts in determining which measures are most applicable to a certain program, especially for unique programs or those that target highly complex populations.

[Slide 50]

When selecting a measurement instrument, program personnel must consider first and foremost whether the measure has established validity and reliability for a given purpose and population. This information may be reported in the relevant research literature or in manuals and websites that accompany a measure. Keep in mind that a short form or alternative version of a given measure is not equivalent to the original version.

Other considerations include training and professional licensure required to administer a measure, which are especially relevant for some diagnostic measures and many tests of cognitive abilities. In addition, the time, costs and usage licenses required to administer and score measures should be considered and balanced with other program priorities. In many cases, alternative measures that require fewer resources are available.

Measures should also align with stakeholders' interests and goals. For example, stakeholders may be more focused on readiness, reintegration, costs or job performance versus symptoms or reducing the incidence of negative outcomes. It is important to measure those areas that are of interest to program personnel and those that are of interest to stakeholders.

Another consideration is how a measure fits with other measures used within a program or within a continuum of services. The goal should be to reduce redundant or overlapping measures within a program while also seeking measures that will allow for comparisons across different points in the service system.

Finally, a measure's practical or clinical utility should be considered. That is, can a measure be used to aid practical or clinical decision-making, such as treatment planning or assessing progress? If so, then the measure has benefits that go beyond program evaluation.

To this point, we have discussed things to consider when selecting a measurement instrument. The other possibility is that you may need to develop a measurement instrument. Note that measurement instruments may be composed of several items. As a result, if you choose to develop your own measurement instrument, you will need to create items, as will be discussed in the slides that follow.

[Slide 51]

Slides 51 and 52 contain a list of 10 evidence-based best practices to consider if program personnel decide to develop customized measures or revise existing measures.

1. First, seek a relatively basic reading level to ensure that the measure is understandable by a wide range of individuals. It is often possible to express even complex concepts using plain language.
2. Second, provide participants with clear instructions on how to complete the measure. Errors often arise when participants don't understand what the evaluator wants.
3. Third, be sure to provide high-quality training and ongoing support and supervision, as well as quality assurance checks to ensure the measure is being used as intended.
4. Fourth, avoid the use of overly complicated wording or questions containing more than one idea. Metaphors and culture-specific language should also be avoided to ensure a measure's content means the same thing to people with differing backgrounds.
5. Fifth, be sure to spell out all acronyms and abbreviations, and only use those that are truly needed. Military environments in particular are saturated with acronyms, and it is important to make sure that all parties know what they actually mean.

[Slide 52]

6. Continuing on slide 52...sixth, ensure the items have face validity, or that they measure what they appear to measure, unless there is a good reason not to do so, such as when you are trying to disguise the underlying intent of the question.
7. Seventh, when possible, be consistent in how you scale the response options. For example, if you use a five-point scale that ranges from very satisfied to very unsatisfied, try to use that same five-point response option scale for all of the items in your measurement instrument. This will ensure clarity when filling out forms and also make sure that scored responses can be summed when necessary.
8. Eighth, it is always best to seek input and feedback on your measurement instrument by examining similar measures or consulting with experts, stakeholders, or program staff and participants. Feedback about items that may be confusing will help clarify those items and reduce error.
9. Ninth, if you need to translate your items from one language into another, consulting with experts is absolutely essential. At a minimum, forward translation from English to the target language and back translation from the target language to English are needed to ensure that a similar meaning is achieved in both languages.

10. Finally, and often overlooked, pilot your measure before using it. This means trying it out on colleagues or an appropriate participant group. Be aware of floor and ceiling effects in which everyone scores very low or very high, respectively. Analyses require that there be some spread or variability in scores in order to be useful.

[Slide 53]

On slides 53 to 55, you will see an example of a poorly written item for a knowledge assessment test, annotated flaws that need to be addressed and a revised version illustrating a better item. The question reads, “Which of the following is the most accurate descriptive phrase representing an individual’s probability of developing PTSD?”

Take a few moments to look at this item before we move on to the next slide. You will notice right away that the question and response choices violate a number of the best practices just mentioned.

[Slide 54]

Slide 54 points out the errors in this item. The item is written using much more complicated language than necessary and contains an acronym that many know but some may not. So, a better question might be phrased, “Which of the following best describes who is most likely to develop posttraumatic stress disorder, or PTSD?”

Now let’s look at the response choices. A is a culture-specific metaphor that could easily be misinterpreted. B is far longer than the other responses and, like the original question, too complicated. C is perhaps a little lengthy, and D is too short and could easily be misinterpreted.

[Slide 55]

Slide 55 shows a revised and much improved item. The question is clearer and spells out posttraumatic stress disorder. The response choices are more similar in length and more straightforward. C, in this case, is the correct answer. Note that this revised item and response choices better assess knowledge that may be gained during training.

[Slide 56]

There are a number of common challenges that arise when determining measurement and evaluation strategies.

[Slide 57]

Certain aspects of military service should always be considered when making important measurement decisions and carrying out data collection efforts.

1. First, some issues that are especially common within military populations may interfere with measurement processes. For example, traumatic brain injuries interfere with memory and concentration; these challenges will affect measurement conducted with individuals who have TBIs. As a consequence, certain adaptations such as taking frequent breaks, using relatively brief tests, or using measures that are more suited to

ability should be considered.

2. Second, in any large organization, wide-ranging abilities and cultural differences must be considered when selecting or developing measures. This is one reason why we emphasize using a sixth to eight grade reading level and avoiding culture-specific language.
3. Third, stigma and concerns about career advancement and clearances are likely to inhibit service members from reporting some issues, especially psychological health and TBI concerns. At the very least, program personnel who conduct measurement activities should be prepared to provide responses to frequently asked questions about how data will be used and who can access data.
4. Last, especially relevant to clinical settings, be sure to measure conditions that frequently co-occur with the presenting problem. Doing so may help a program to identify additional benefits beyond its primary focus and will also serve to clarify the needs of the target population.

[Slide 58]

In the next several slides, we have prepared a brief FAQ on common challenges based on questions we have received when discussing evaluation and measurement strategies with program personnel.

[Slide 59]

On slide 59, “If my prevention program is successful, how do I measure something that did not happen as a result?”

This is a question that occurs quite often, and it should be noted that a large proportion of our audience today comes from non-clinical programs, making this question very relevant and timely. Prevention is not just about reducing the long-term incidence of problems or risk for problems. According to the Institute of Medicine, it is also about health promotion.

Non-clinical programs should have both short-term and longer-term goals relevant to health promotion and risk reduction, and conduct measurement activities accordingly. In the short-term, measured processes may include the number of participants, or perhaps referrals made as a result of participation. Short-term outcomes that relate to longer term objectives for decreasing risky behaviors and increasing health-related behaviors might include learning or enhanced awareness resulting from participation in the program.

There are a number of service-level databases that track long-term outcomes, and in some cases it may be possible to link short-term outcomes to long-term outcomes. Regardless, program personnel should be responsible for measuring short-term outcomes – those that occur during the program, at program end, and perhaps over the first few months after participants have completed the program.

[Slide 60]

On slide 60, another frequent comment is, “I cannot find measures with demonstrated validity and reliability for the purposes and population of my program.”

In the information age, a common problem is that there is so much information it is difficult to know where to begin to find the information you really want.

1. The best starting place is often to consult experts within your program area, such as researchers or consultants. In most cases, people in these roles will be pleased to help out and are relatively easy to reach.
2. Second, it is likely that measures have already been developed for purposes or populations similar to those of your program. Pilot testing and conducting focus groups can help you determine whether an existing measure can really be used in your setting.
3. Third, make use of the best practices for item development discussed earlier in the presentation. It is often possible to adapt an existing measure using these practices for item development. Keep in mind if you adapt existing measures, you will also need to re-examine validity and reliability within the program.
4. Lastly, as we have mentioned, if there really are no high-quality existing measures, you may need to develop a new measure. This may be especially necessary when assessing highly specific target outcomes such as learning related to program participation. Again, make use of best practices in item development.

[Slide 61]

Finally, on slide 61, one more very common challenge expressed by program managers is, “My program lacks the resources, such as time, training, materials and funding, to conduct measurement activities.”

This is a very common concern for new programs and those considering making changes in their measurement practices. The basic issue is that program personnel think they cannot afford all of the resources that it will take to engage in high-quality measurement. We might pose the question, “Can you really afford not to measure inputs, activities, outputs and outcomes, or to do poor-quality measurement on these components?”

Conducting measurement is an important investment in the program’s future. Measurement informs program leadership about what is working, what is not working, and what should be improved. Moreover, providing data about what is working ensures a program’s survival in this age of accountability.

In addition, it may be possible to examine how time and other resources are currently being used and to find areas in which some processes could be eliminated to make more room for measurement activities and the ongoing support and quality assurance they require.

Lastly, many measurement materials are free or low-cost, as are many of the consultations and training activities needed to measure effectively. Costs, therefore, may be a more surmountable barrier to effective measurement than you might initially think.

Now, Captain Thoumaian will conclude the webinar and highlight a few key takeaways.

[Slide 62]

CAPT Thoumaian: Thank you Dr. Best, Dr. Sawyer, Ms. Aguirre.

You've heard a great deal today about designing program evaluation and measurement strategies, which are critical in program evaluation efforts. Choosing the right strategies as part of the evaluation planning process is an important aspect of any high quality evaluation effort.

[Slide 63]

A key takeaway is that decisions about evaluation designs should be based on multiple considerations, including evaluation goals, the nature and intent of a program, and a program's level of maturity. By taking these factors into account, evaluators can ensure that an evaluation effort results in useful information that can serve the needs of all those involved with a program and in particular the service members and families that participate in them.

A second key takeaway is that programs can use evaluation data generated to firmly establish connections between a program's objectives, and its inputs, activities, outputs and outcomes.

Finally, careful planning and attention to the accuracy of measurement ensures that the data generated by measurement activities truly reflect what a program is doing, and whether it is achieving its objectives. Only then can a program focus in on ways to improve the program's quality and effectiveness.

Evaluation should be an essential part of everyday program operations. Once evaluation activities are built into the fabric of a program, it will be easy to extract information to demonstrate accountability and to enhance a program's ability to serve its participants.

I hope you will continue to attend these training presentations and also consult the Program Evaluation Guide and other resource materials on the DCoE website.

Now back to Ms. Aguirre.

[Slides 64 through 68]

Ms. Aguirre: Thank you, Captain Thoumaian. There is a great deal of useful information available to programs on program evaluation. On slides 65 through 68, we provide a list of relevant references and resources that we think may be useful. These include a number of resources to help you identify specific measures and metrics relevant to your program.

[Slides 69 through 74 are provided for reference]

[END]