



Strategies for Measuring Program Processes and Outcomes

Presented on May 20, 2014

Episode 3 in the Program Evaluation and Improvement Training Series

Presenters

CAPT Armen Thoumaian, Ph.D.

Health Science Officer

Office of Policy, Programs and Integration

Defense Centers of Excellence for Psychological Health and Traumatic Brain Injury (DCoE)

Silver Spring, Md.

Aaron Sawyer, Ph.D.

Research Scientist

Contract support for DCoE

Arlington, Va.

Patrick High, Dr.P.H.

Epidemiologist

Contract support for DCoE

Arlington, Va.

Moderator

Debra Stark, M.B.A.

Research Scientist

Contract support for DCoE

Arlington, Va.

[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 another episode in the 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 Centers: the Defense and Veterans Brain Injury Center (or DVBIC), the Deployment Health Clinical Center (or DHCC), and the National Center for Telehealth and Technology (or T2). DCoE and its Centers work closely with one another to promote high-quality prevention and care across the Defense Department.

Together, we produce a variety of trainings on subjects ranging from program evaluation to clinical care and prevention practices.

The DCoE Program Evaluation and Improvement training series is designed to increase the capacity of psychological health and traumatic brain injury programs to engage in program evaluation activities.

The trainings in this series are directed toward program administrators and service leadership who are currently involved with or plan to conduct program evaluation activities.

This series contributes to DCoE's larger mission to enhance the quality and effectiveness of psychological health and traumatic brain injury programs by providing training on key activities that may be used to advance program evaluation and improvement efforts.

On behalf of the Defense Centers of Excellence for Psychological Health and Traumatic Brain Injury, thank you for participating in this training series.

[Slide 1]

Ms. Stark: Hello. My name is Debra Stark. I provide contract support to the Defense Centers of Excellence, and I will be your moderator for this presentation, Episode 3 in the program evaluation and improvement 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 "Strategies for Measuring Program Processes and Outcomes." 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 located on the left of your screen. You can do this at any time, 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 yourself 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 as the connection catches up with the audio at times. Depending on your network security settings, there may also be some noticeable buffering delays.

Closed captioning is not available for this event.

[Slide 3]

Continuing education credit is not available for this event but may be available for future webinars. Webinar materials from this series will be made available in the Program Evaluation section of the DCoE website. For information about other DCoE webinars and trainings, visit the

Training section of the DCoE website by following the link on slide 3. Slides and other materials are available in the boxes at the bottom of the screen during the webinar.

[Slide 4]

This webinar was introduced by Captain Armen Thoumaian. Captain Thoumaian is a Health Science Officer with 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 completed a National Institute of Mental Health fellowship in Community Mental Health.

[Slide 5]

Our first presenter is Dr. Aaron Sawyer, a Research Scientist providing contract support to DCoE. Dr. Sawyer 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 an M.S. in Experimental Psychology and a Ph.D. in Clinical Psychology. He completed postdoctoral training at The Kennedy Krieger Institute/Johns Hopkins University and is a Licensed Psychologist.

Our next presenter is Dr. Patrick High, an epidemiologist providing contract support to DCoE. He has over a decade of experience and has expertise in survey design, research methodology and program evaluation. His experience includes supporting the Office of the Under Secretary of Defense for Personnel and Readiness, Operations Research and Safety and the Defense Suicide Prevention Office as an epidemiologist. Dr. High holds the degree of doctor of public health with specialization in Epidemiology and Biostatistics from the Uniformed Services University of the Health Sciences and previously spent nine years in the Illinois Army National Guard.

[Slide 6]

Dr. Sawyer: Our final presenter is Ms. Debra Stark, also a Research Scientist providing contract support to DCoE. Today, Ms. Stark is both moderator and presenter. She is a survey methodologist and analyst with more than 15 years of research experience. Ms. Stark's work includes program evaluation and monitoring, qualitative data analysis, and survey instrument design. She has worked on health services evaluation projects with several Federal agencies, including the Department of Veterans Affairs and Tricare Management Activity. Ms. Stark received her MBA from Vanderbilt University.

[Slide 7]

Ms. Stark: This presentation will provide guidance on metrics and measurement strategies for program managers and others who seek to measure processes and outcomes during the course of program evaluations. The content in this presentation is intended to apply to a wide range of psychological health and traumatic brain injury programs. Training on highly specific measurement tools for a particular program are beyond the scope of this presentation; however, we do provide information regarding where to find additional resources and measurement tools specific to individual programs.

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

- Explain how metrics and measurement strategies are used in program evaluation and improvement efforts
- Demonstrate knowledge of important considerations for selecting or developing measures
- Implement suggested guidance to collect and store evaluation data, and
- Identify common challenges that programs face in measuring processes and outcomes

[Slide 8]

As seen on slide 8, the first topic is an introduction to metrics and measurement, followed by an examination of measurement domains, best practices in measure selection and development, data collection and storage, and common measurement challenges.

Dr. Sawyer will begin the presentation, followed by Dr. High and myself with concluding comments from Captain Thoumaian.

[Slide 9]

Dr. Sawyer: I am proud to be here today as a researcher and psychologist supporting DCoE's and the military's efforts to provide high quality prevention and care for service members and their families.

Trainings on measurement strategies and other program evaluation topics support a culture of effectiveness in which programs are well-defined, generate accurate data and are implemented effectively.

[Slide 10]

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 cruising around islands in the Pacific Northwest with his father creating intelligence test questions.

Around the time I studied with him, Robert M. Thorndike was doing some large-scale landscaping at his house. He suspected that the stone yard that provided his landscaping rocks had 'shorted' him on his 2-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 trucking out the missing stones and 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. This training episode is about how programs produce data on their processes and outcomes through measurement.

[Slide 11]

Before we get too far along, it is necessary to clarify a few key terms. When we use the term 'measurement,' we are referring to the process of collecting data, or information about some matter. A 'metric,' on the other hand refers to a specific type of data; that is, units of data that may serve as indicators of the activities of a program's staff or its participants, as well as the performance of a program in meeting its objectives. A 'measure' is a tool used to collect data, such as a test, questionnaire, survey, or instrument.

[Slide 12]

The U.S. Military has been involved in the measurement business for quite some time, beginning in World War I with the Army Alpha and Beta aptitude test batteries, which assessed the skills and abilities of service members.

The Defense Department collects, stores and assesses a great deal of data on service members through such entities as the Armed Forces Health Surveillance Center, Defense Manpower and Data Center, DoD Suicide Event Report database and many other efforts. Essentially, being a part of the military means being measured on a regular basis.

In addition, the military, like other large organizations, relies on measurement strategies to capture and analyze data for program evaluation and improvement efforts, as well as guide the selection, development, and funding of the full range of military psychological health and traumatic brain injury programs.

[Slide 13]

Of course, measurement is not limited to military life. Think for a moment about all the ways in which you may be measured during the course of a week.

When you drive, information about your speed is measured by your speedometer, or perhaps by a speed camera on a highway construction site. When you surf the web, your location is recorded, as are social media posts. Likewise, you are measured on minutes of phone usage by your cellular provider, and information about your productivity and attendance is documented at your workplace. Information about your state of health is collected by your physician, and federal and state agencies regularly collect information about your age, income and travel outside the country.

[Slide 14]

Let's review some important concepts underlying any type of measurement of program processes or outcomes:

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 using a questionnaire given to a service member returning from deployment. I give the service member the form and get a score. That score is related to the actual risk but also contains some error, perhaps because the service member had difficulty concentrating, or the questions were

difficult to understand, or perhaps the service member did not want to answer affirmatively for items 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 measurement is to minimize error so that measured data points are as close to the true value as possible.

[Slide 15]

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 estimate a person's risk for 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 16]

Today, our focus is on quantitative data, which is information expressed in numerical form. There are four basic types of quantitative 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 1 for women, but the difference is a matter of type rather than women being 1 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 1 and 2 isn't necessarily the same size as the difference between a rank of 2 and 3. 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 1 and 2 is the same amount as the difference between 2 and 3. 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 more important when we discuss analyzing and interpreting data in the next episode in this training series.

[Slide 17]

Qualitative data are often considered to be "soft" data. They are often quite valid in terms of accurately representing characteristics or situations; however, many consider qualitative data to

be less reliable than quantitative, or “hard”, data since they are often open to interpretation and difficult to compare directly to other types of information. Regardless, qualitative data are still important and widely-used.

Some of the most commonly-used qualitative data types are listed here.

- In-depth 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.
- Focus groups use a leader to guide a structured discussion among 4 to 9 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 characteristics like status or rank.
- 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 where staff members who participated in an activity discuss its strengths and weaknesses.
- 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.

Qualitative and quantitative data are not necessarily competing forms of data. It is possible and often advisable to use multiple forms of data to monitor program functioning and results. There is no one measurement tool that is best in all cases and the choice of which data collection strategy to follow depends on many factors, as we discuss throughout this presentation.

[Slide 18]

Your program’s stated objectives should influence the development of specific metrics that reflect program processes and outcomes. Metrics produce data about those processes and outcomes, which can then be analyzed to determine whether objectives are accomplished. This connection between objectives and the use of measurement strategies to determine whether they are achieved is essential to program evaluation and improvement efforts, both within programs and in regards to evaluations conducted by external evaluators.

Now, I will hand off the presentation to Dr. High.

[Slide 19]

Dr. High: Thank you Dr. Sawyer. I will briefly recap some of the messages from Episode 2 in the series to provide context for choosing measurement domains.

[Slide 20]

In Episode 2, we said that a mission statement explains the purpose for a program's existence. Mission statements encompass everything the program does and intends to achieve, and should align with the mission of the broader organization in which the program exists.

Goal statements outline in more detail what a program intends to accomplish.

Objectives are the most specific type of program planning statement. They describe goals, or even parts of goals, in terms of highly specific units that can be measured to determine whether a program is working. We recommended that objectives be SMART: specific, measurable, achievable, relevant, and time-bound.

Objectives may relate to the processes an organization uses, such as holding a certain number of program sessions or activities, or objectives may relate to the desired outcomes of the program.

In this webinar episode, we are looking into the "M" part of SMART, that is "measurable" objectives and how to select measurement domains and more specific measurement strategies within those domains.

[Slide 21]

How will we, or anyone else, know that expected processes or outcomes have occurred? Consider what evidence is needed to show that program activities are occurring, or that participants have exhibited changes. This evidence must be collected and analyzed in order to compare the results with stated objectives.

In the example on slide 21, consider a web-based training on traumatic brain injury, or TBI, for unit commanders. That stated objective is that program staff will provide 50 training sessions to unit commanders, who will demonstrate increases in their knowledge of TBI symptoms.

Under "Measured Process," we see that 35 web trainings took place. Clearly, 35 trainings are fewer than the objective to provide 50 trainings. So, this measured process is evidence that the objective was not met. This may have implications for improvements to increase the number of trainings, or it may indicate that the initial objective was unrealistic. Either way, the comparison is highly informative and could not occur without a clear objective in place and subsequent measurement to determine whether the objective is met.

Under "Measured Outcome," we see that unit commanders, on average, increased their knowledge of TBI symptoms and management strategies by 20 percent. The measured outcome, in this case, appears consistent with the stated objective, so you could say that the objective has been met.

The knowledge gained by such measurement and comparison processes is invaluable in the effort to maintain or improve quality in a program.

[Slide 22]

Slide 22 depicts the basic logic model form presented in Episode 2.

For a quick review, inputs are what the program needs, and includes the resources a program requires. Outputs are what the program does, and includes the activities and products of a program. Outcomes are the changes that result from a program.

With this in mind, use your program's logic model to develop measurement questions and metrics for the program. As evaluators, we will want to know: What activities are being conducted? Are the right activities occurring? Are they being conducted correctly? Are improvements or changes needed?

Likewise, we want to know what outcomes are occurring in the short-term and over greater lengths of time. Are the desired outcomes occurring? If so, what is the size and scope of the outcomes? Are any unintended or adverse outcomes occurring?

The logic model can help guide the selection of measurement domain choices, depending on whether we want to measure inputs, outputs, or outcomes. Domains are the topic areas of interest that we want to be sure to include and measure.

The next few slides discuss domains that correspond to each area of the logic model.

[Slide 23]

Inputs are the things that are needed to implement the program or intervention. Track these items, using forms, logs, recordings, or other material.

Note the demographic composition of program staff. It may turn out that there is an association between the demographics of program staff and the demographics of program participants, but we'll never know if we don't track it.

Track the training sessions that staff undergo within your program. Note whether the training is ongoing, or if it is a one-time activity. Note the supplies needed for the program, including office space and equipment. This will be very important information to have when it comes time to determine program costs. Be sure that you have metrics and processes to track these inputs.

[Slide 24]

Track all program delivery activities. For example, if one activity a program lists is providing classes on reintegration for service members, the output for this activity would be the number of classes staff members deliver per week, per month or per year, and the number of attendees per class and overall. An area often overlooked is program promotion and outreach activities that staff conduct. These are outputs, too. See the activities listed on slide 24. Every time your personnel sit in a booth at a health fair, take part in a radio interview, make follow-up calls to program participants, or tack up posters, those activities should be captured and counted. Did a staff member speak or present at a conference or on a panel? Retain a copy of the presentation, conference proceedings, and conference agenda? Also, maintain meeting minutes and records of action items?

[Slide 25]

Track information about program participants and their attendance and involvement in program activities. Track demographic information about participants: in addition to the standard

questions about age, gender and ethnicity, note their service branch, rank, duty status, and deployment status. Follow up on your participants, if possible. Did they actually access the referral that the provider made, or fill that prescription? You will need Social Security numbers for this level of data collection effort, but if you have that, you can follow on in TRICARE, or access information from Defense Manpower Data Center .

Record, catalogue and store information about your program. Describe program accessibility, eligibility criteria, wait list time and so on.

Keep statistics on program activities: the number of pamphlets distributed, the number of persons who sign up for a webcast, or downloads from a website. Whatever your unit of service or activity is, be sure that at least one of your metrics tracks these.

[Slide 26]

What every evaluator wants to do is connect outputs—what the program does—with outcomes—the change the program makes in someone’s psychological or physical health and well-being. Tracking outputs and outcomes provides the evidence as to whether program activities produced the desired results.

You may be wondering why the word “facts” is in quotation marks on this slide. This is because you will not be conducting a research study—but rather an evaluation study. Although the association between a certain activity and a certain outcome may seem like a cause-and-effect relationship, it is unlikely that you will be conducting a formal research trial in a way that would allow you to make that determination. However, measurement can indeed help to make the case that program processes are related to desired outcomes. Measurement can also help you to examine programmatic trends over time, and help you to decide whether any adjustments or changes made to the program result in improvements.

For example, as evaluators, we may want to know if people with post-traumatic stress disorder experience a decrease in symptom levels after completing an outpatient treatment program. Process metrics like the number of sessions provided can be linked to target outcomes to determine if there is a relationship between the service provided and the participant’s outcome.

[Slide 27]

We have highlighted how important it is to track outcomes, those changes we hope program participants will experience as a result of their participation in a program. You may be asking, which measurement domains should we track?

You should seek to measure multiple domains to ensure that all relevant outcomes are captured. Listed on slide 27 are example domains that are likely relevant to each of your programs. These domains are outcome areas that program staff should measure and about which evaluators and stakeholders will likely want to know.

[Slide 28]

Specific metrics are needed to provide data about outcomes within each domain or topic area. For instance, under the “Resiliency” domain, which we saw on the previous slide, our metrics or measurement instruments should produce information on number of subdomains or specific outcome areas.

In this case, for resilience, we would measure a participant's coping and problem solving abilities, self-esteem, and social support. We would not measure personality type as part of this domain. Do not collect data from more metrics or domains than are necessary, but do measure pre- and post-intervention to determine whether changes have occurred relative to a baseline.

The next few slides examine output and outcome measures for a hypothetical domestic violence program.

[Slide 29]

In terms of output, or process measures, we have stated that it is important to track activities, participation and products.

Consider what is important to measure for a domestic violence program. Many people know to count the number of group sessions held, but what about counting other, less obvious activities? When a participant is referred to a housing coordinator or when program staff provide outreach or education services, those are output or process measures that need to be captured by measurement activities as well. They provide important information about program functioning and level of effort. Likewise, program satisfaction is an important domain in terms of engaging participants and generating future participation.

Again, regarding participation, perhaps tracking the number of calls that a Helpline receives is obvious. However, note the less obvious, too--as when people external to the domestic violence program refer someone to it. Track how many external referrals occur and their sources. This may lead to important improvements or program modifications in the future.

And of course, program staff should track all products created by the program: whether a website was developed and the number of hits it received or whether posters were produced and how many or where they were distributed.

[Slide 30]

What might program staff want to track related to outcomes? We displayed several domains back on slide 27 that would be appropriate here, and you can see them in the bullets on slide 30—Awareness, Quality of Life and Family Relationships. Next to each bullet are the relevant metrics. First, under Awareness, program developers and staff hope that after participating in a certain number of sessions, participants would experience a shift in their attitude toward domestic violence, be able to recognize risk behaviors, and know about support and legal services available to them. Second, changes in quality of life that may result from program participation could be measured instruments that assess stress level, quality of life or perhaps perceptions about the future. Finally, the quality of family relationships is important to measure for this type of program; this domain could be assessed by looking at incidents of domestic violence, parent-child conflict ratings, or communication quality.

Measures for this program should ideally be collected at multiple time-points, such as before the domestic violence program is provided, three months into the program, upon completion and possibly at follow-up.

[Slide 31]

Ms. Stark: Thank you, Dr. High. The next portion of the webinar will provide guidance on selecting or developing measures for program processes and outcomes.

Measurement strategies are most effective when they use multiple methods and multiple 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, 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 32]

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 objectives and population. Generally speaking it is best to choose existing measures with proven reliability and validity. There are a number of resources that can serve as aides to measure selection for specific programs, including those listed on slide 55, such as the Agency for Healthcare Research and Quality, the National Quality Forum, 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 novel programs or those that target highly complex populations.

[Slide 33]

If selecting an instrument, program personnel must consider first and foremost whether the measure has established validity and reliability for the program population. This information may be reported in the relevant research literature or in manuals and on websites that accompany the measure.

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 stakeholder's interests and goals. For example, stakeholders may be relatively more focused on readiness, reintegration, costs or job performance versus symptoms or reducing the incidence of negative outcomes. It is necessary to measure processes and outcomes that are both of interest to program personnel and 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 it holds benefit beyond program evaluation.

[Slide 34]

Slides 34 and 35 contain a number of research-informed best practices to consider if program personnel decide to develop customized measures or revise existing measures.

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.

Plain language also means avoiding the use of overly complicated wording or questions containing more than one premise. Metaphors and culture-specific language should also be avoided to ensure a measure's content means the same thing to people with differing backgrounds.

Third, unless there is a good reason not to do so, ensure items have face validity. That is, items should measure what they appear to measure.

Fourth, when possible, be consistent in the scaling used for responses. This will ensure clarity when filling out forms and also make sure that items can be added when necessary.

Fifth, it is always best to seek out input and feedback on measurement by examining similar measures, consulting with experts, stakeholders and staff, and by testing the items with program participants to get direct feedback about items that may be confusing.

[Slide 35]

Continuing from the previous list, it is best to avoid altering or shortening existing measures. A shortened or altered version of a valid and reliable measure is not necessarily as valid and reliable as the original.

In the event item translation is needed, consulting with experts is absolutely essential. At a minimum, forward translation from English to the target language and back translation from the target language back to English are needed to ensure that similar meaning is achieved in both languages.

Finally, and often overlooked, pilot a measure before use. 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. Analyses require that there be some spread or variability in scores in order to be useful.

[Slide 36]

On slide 36, you will see an example of a poorly written item. The question reads, "Which of the following is the most accurate descriptive phrase representing an individual's probability of developing PTSD?"

You will notice right away that the question and responses violate a number of the best practices just mentioned.

[Slide 37]

Slide 37 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 how likely it is that someone will develop posttraumatic stress disorder, or PTSD?”

Then come 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 just right, and D is too short and could easily be misinterpreted.

[Slide 38]

Of course, determining which metrics or measures to use is only half the battle. Next, program personnel must actually collect the information and store it in a way it can be easily accessed for analyses, the subject of an upcoming program evaluation training series episode.

[Slide 39]

To begin the planning process for collecting and storing data, develop a data matrix. Slides 39 and 40 present examples of data matrices, the first for a clinical program and the second for a non-clinical program.

In the first example, consider a clinical program focused on mental health outcomes. In this program, three different metrics are used to measure the effects of the program on participant outcomes. In the Metric 1 column, providers collect questionnaires from program participants before treatment, after treatment 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 treatment planning and outcome tracking.

In the Metric 2 column, providers give participants a functional rating before, after and 3 months following treatment. The provider is assigned responsibility for inputting data, and the data will be used for outcome tracking purposes and to develop reports to stakeholders.

The third metric also involves functional ratings, but in this case, providers collect information from unit commanders at pre- and post-treatment. They then provide the information to program managers to input and analyze for outcome tracking and reporting purposes. In many ways, this outcome matrix is exemplary, because it involves multiple methods of measuring the same outcome domain and uses different sources of information.

[Slide 40]

On slide 40, we provide a second data matrix example, this time for a non-clinical web-based training program. In this case, there are two process metrics assessing participation and one outcome measure assessing learning resulting from participation.

In the first participation metric column, participation is assessed by the webinar provider using the number of attendees on the day of the webinar. The webinar event provider then enters the data, and all of the data are analyzed by the program manager. This information is subsequently used to track outputs and inform reports. The second participation metric column is a different way to assess participation using website downloads and page visits or “hits.” This information is collected monthly by the website manager.

In the last column, learning outcomes are assessed using a post-test given by the provider as part of the web interface, much as we ask you to provide feedback for this webinar or as a learning quiz would be administered as part of continuing education credits. This information may be used for multiple purposes, such as outcome tracking, reports to stakeholders and award of continuing education credits to attendees.

[Slide 41]

A data collection and storage plan can be easily constructed from the initial foundation of data matrices. This plan should be comprehensive, covering how all data on inputs, outputs and outcomes are collected on an ongoing basis. In addition to clarifying the information in the data matrix, a data collection and storage plan should also provide guidance regarding how data cleaning and quality assurance processes are to be conducted. Recall from the beginning of the presentation that an important measurement goal is to minimize error – in fact, it is likely that at least as much error comes from data mismanagement as from any limitations of the measure, itself. Consequently, it is extremely important that to carefully review that data to ensure that data are complete and that they are entered accurately.

[Slide 42]

A data collection and storage plan will also lead program managers and staff to consider the inputs needed to effectively conduct measurement activities. Measurement can often be time consuming, especially when introducing a new process. Staff will also need appropriate training to complete all relevant procedures, and programs may need additional funding for materials. Not to be overlooked, it is absolutely essential that proper input support be given to quality assurance and maintenance tasks.

[Slide 43]

In all data collection and storage procedures, staff should seek to protect participants’ privacy and confidentiality within the bounds of regulatory and ethical requirements.

Sensitive data include personal identifiable information, or PII, which refers to information that can be used to identify an individual participant, and protected health information, or PHI, which includes information about health status. Protections for these types of information may fall within regulations pertaining to the Health Insurance Portability and Accountability Act, or HIPAA – it is best to consult applicable policies, stakeholders and program administration to determine which regulations apply.

[Slide 44]

Regardless of any particular regulations, there are a number of best practices that apply to protecting participant data, including use of secure data storage in locked rooms and only in password-protected computer systems. It is also best when possible to use ID numbers to separate participants’ identifying information from any type of health or outcome data and to always use secure methods of transmitting data. Procedures must be clarified using standard

operating procedures, or SOPs for handling data, and appropriate staff training must be provided with ongoing support. In rare cases, it may be possible to use anonymous data collection; however, it is generally advisable not to use anonymous data, because doing so makes it impossible to match participant information to key outcomes and processes, or to track participants over time

[Slide 45]

Before we end this section, I will leave you with a brief list of ways to minimize risk when collecting and storing data. To highlight a few examples, it is best to limit or prohibit altogether any off-site handling of data and to really make sure on-site paper and electronic data are secure. It is also advisable to track adverse events, so that program personnel can learn from them, and it is best to collect only the data a program needs.

[Slide 46]

Dr. Sawyer: Thank you Ms. Stark. There are a number of common challenges that arise when conducting measurement activities.

[Slide 47]

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

First, some issues that are especially common within military populations interfere with measurement processes. For example, traumatic brain injuries interfere with memory and concentration; these challenges will affect any type of 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.

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 6th to 8th grade reading level and avoiding culture-specific language.

Third, stigma and concerns about career trajectories and clearances are likely to inhibit service members from reporting issues of any kind and psychological health and TBI concerns in particular. At the very least, program personnel who conduct measurement activities should be prepared to provide responses to frequently answered questions about how data will be used and who will be able to access data. As an example, we provide a link on slide 55 to the Real Warriors Campaign guidance on answering Question 21 of the "Questionnaire for National Security Positions," also known as Standard Form 86.

Last, especially relevant to clinical settings, be sure to measure conditions that frequently co-occur with the target 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 48]

In addition to military-specific issues that impact measurement, there are several more general factors that threaten the validity and reliability of measurements.

Inconsistent procedures for measurement can make the results incomparable, so it is important to standardize measurement processes. Likewise, attrition, which refers to the loss of participants for follow-up, and incomplete data collection, will make a program's data less representative of the target population.

In the third bullet on slide 48, you may note we have said this about half a dozen times, so you know it is very important. Using a measure for purposes or populations for which it was not designed makes the results difficult to interpret as they don't mean the same thing in a different context. For example, using a deployment-focused TBI risk assessment that asks about exposure to combat and blasts would not make sense for a non-deployed population or for service members who are just entering the military.

Finally, program staff members should remain neutral during measurement processes and avoid conveying information that might bias results. This is especially true in situations where performance is being assessed or in clinical settings where participants may have close relationships with the program staff member.

[Slide 49]

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

[Slide 50]

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

This is a question we hear all the time, and it should be noted that a large proportion of our audience 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 might include learning or awareness resulting from participation or changes in risk behaviors and health-related behaviors related to long-term outcomes.

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, personnel at the program-level 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 complete the program.

[Slide 51]

Another frequent comment is: “I cannot find measures that have been shown to be reliable and valid 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.

The best starting place is often to consult experts within your program area, such as researchers, consultants or similar programs. In most cases, people in these roles will be pleased to help out and are relatively easy to reach.

Second, it is likely that measures have been developed for purposes or populations similar to those of your program, and pilot testing and focus groups are likely to help determine whether a measure can really be used in your setting.

Third, make use of the best practices for item development discussed on slides 34 and 35. It is often possible to adapt an existing measure using these practices for item development and then re-examine validity and reliability within the program.

Lastly, as we have mentioned, if there really are no high-quality existing measures, you may develop a new measure, especially when assessing learning related to program participation or highly specific target outcomes. Again, make use of best practices in item development.

[Slide 52]

One more very common challenge expressed by program managers is: “My staff lack the resources, such as time, training, materials and funding, to carry out high-quality measurement.”

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 or to do poor-quality measurement?”

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 the age of accountability.

In addition, it may be possible to measure 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 perceive.

[Slide 53]

CAPT Thoumaian: Thank you Ms. Stark, Dr. High, Dr. Sawyer.

You’ve heard a great deal today about measurement of program processes and outcomes, which are a critical part of building a culture of effectiveness in the Defense Department’s system of care for psychological health and traumatic brain injuries.

[Slide 54]

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

Without measurement, objectives are merely aspirational statements that cannot be used to improve program quality and effectiveness.

Careful planning and attention to measurement quality ensures that the data generated by measurement activities accurately reflect what a program is doing and whether it is achieving its objectives.

Measurement is an essential part of program operations and should be an everyday activity. Documenting program processes and outcomes will enable program managers to improve quality and outcomes for service members and their families.

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.

[Slide 55]

Ms. Stark: Thank you Captain Thoumaian. There is a great deal of useful information available to programs on measurement tools such as surveys and checklists, and on data privacy and storage considerations. On slide 55, we provide a brief list of key references and resources that we think may be useful.