UCLA CRESST NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

APPROACHES FOR COMBINING COMPETENCIES AND METRICS AS A SINGLE WEIGHTED METRIC

John J. Lee, Alan D. Koenig, Kilchan Choi, Michael Senteney, Eva L. Baker, Joanne K. Michiuye, and Harold F. O'Neil

NOVEMBER 2024

CRESST REPORT **873**

This work was supported by contract number 24361822D00001 24322623F0104 (PO100238) from the Office of Personnel Management via PowerTrain Inc. with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles.

To cite from this report, please use the following as your APA 7th edition reference: Lee, J. J., Koenig, A. D., Choi, K., Senteney, M., Baker, E. L., Michiuye, J. K., & O'Neil, H. F. (2024). *Approaches for combining competencies and metrics as a single weighted metric* (CRESST Report 873). UCLA/CRESST.

Table of Contents

Approaches for Combining Competencies and Metrics as a Single Weighted Metric

John J. Lee,[a] Alan D. Koenig,[a] Kilchan Choi,[a] Michael Senteney,[a] Eva L. Baker,[a] Joanne K. Michiuye,[a] and Harold F. O'Neil [b]

[a] CRESST/University of California, Los Angeles
[b] USC Rossier School of Education

Abstract: This report discusses developing a single composite metric or score from various U.S. Navy training programs to allow for comparisons of training effectiveness across programs. It describes incorporating three levels of the Kirkpatrick model currently used to evaluate training: reaction, learning, and behavior. The goal is to create a composite score combining Levels 1, 2, and 3 that can serve as a key performance indicator for evaluation purposes. Several weighting methodologies are proposed. The document outlines various Navy data sources that could be included in the composite score, such as (Personnel Qualification Standard (PQS) time to qualify, Surface Training Readiness Management System (STRMS) assessment scores, advancement exams, and baseline questionnaires. Validity and reliability of the composite score are important considerations that expert weighting of the measures may help establish. An example approach is described using CRESST data to illustrate the qualitative and quantitative weighting methodologies.

Introduction

The Naval Training and Education Command (NETC) develops and implements U.S. Navy training. NETC is responsible for managing the training, including the instructors, and looking at the effectiveness of the training based on the Kirkpatrick model of evaluation (Kirkpatrick & Kirkpatrick, 2016, defined below). Most Sailors first undergo recruit training and then proceed to "A" schools, responsible for introducing and familiarizing Sailors with tasks performed for a given enlisted rating (or job). There are over 90 enlisted ratings (non-officer positions) ranging across various fields, such as combat systems, engineering, culinary, munitions handling, and medical, to name a few. Depending on the rating, additional instruction may be taken through "C" Schools, which allow the Sailor to specialize in a particular system used on the ship.

For stakeholders, primarily higher up in the chain of command in the Navy, numbers and expert human judgment provide real-time or near real-time information about the status or readiness of commands and their associated subcommands at varying levels of aggregation. These data can be used for high-level management and policy decisions, and using composite scores would benefit decision-makers with at-a-glance metrics combining competencies and metrics as a single weighted metric for evidence-based decision-making. That information, however, will only be helpful if this single metric is accurate and timely, and the measures, metrics, and indicators used are of good technical quality, that is, have sufficient validity and reliability. With advancements in technology, capturing and accessing the requisite data are becoming more feasible. Still, some security policy and infrastructure hurdles are necessary to document the required permissions and to protect classified information. Most data are not classified in the MyNavy Learning / Surface Training Readiness Management System (MNL/STRMS) systems. These nonclassified data are the focus of this discussion.

This report aims to describe several possible approaches for deriving a composite score for competencies that reflect and incorporate Levels 1, 2, and 3 of the Kirkpatrick model (Kirkpatrick & Kirkpatrick, 2016). The Navy uses this model to evaluate training effectiveness across the learning continuum from "A" Schools to on-the-job training. This report details the data currently collected, followed by an analysis of possible future data sources, and finally, describes with a detailed example how the data could be used to create a composite.

Overview of Kirkpatrick Levels and Data Sources

Table 1 provides definitions of the four levels in the Kirkpatrick model, reprinted from Kirkpatrick and Kirkpatrick (2016, p. 27). CRESST does not have access to any Navy data related to Kirkpatrick Level 4, and therefore only Levels 1, 2, and 3 will be discussed in this report.

Table 1
Kirkpatrick Model of Training Evaluation

Kirkpatrick level	Definition
Level 1: Reaction	The degree to which participants find the training favorable, engaging, and relevant to their jobs
Level 2: Learning	The degree to which participants acquire the intended knowledge, skills, attitude, confidence, and commitment based on their participation in the training
Level 3: Behavior	The degree to which participants apply what they learned during training when they are back on the job
Level 4: Results	The degree to which targeted outcomes occur as a result of the training and the support and accountability package

Currently Collected Navy Data

Level 1: Reaction

As prescribed in NAVINST1540-2C (U.S. Navy, 2023a), Level 1 comprises end-of-course questionnaires given to students. Course content in terms of scope and depth is evaluated with the following six questions:

1. This training met my needs.
2. Course content enabled me to meet the training objectives.
3. The physical training environment (classroom, labs, etc.) enabled me to meet the training objectives.
4. The sequence of the lesson topics helped me to meet the training objectives.
5. Course and support materials helped me to meet the training objectives (e.g., trainee guides/handbooks).
6. Training aids were used effectively to help me meet the training objectives (e.g., videos, mock-ups, etc.).

Comments are also solicited for each question. All questions use a Likert-type scale with five items as options: *strongly disagree, disagree, neither agree nor disagree, agree,* and *strongly agree*.

In addition, eight questions ask about the quality of the instructors (primary and if applicable, secondary) using the same scale. Comments are also solicited for these questions. All questions use a Likert scale with five options: *strongly disagree, disagree, neither agree nor disagree, agree,* and *strongly agree*.

These questions are the same for primary and secondary instructors and include:

1. The instructor was professional at all times.
2. The instructor kept the learning interesting.
3. The instructor was prepared.
4. The instructor taught at a level I could understand.
5. The instructor clearly explained the training objectives.
6. The instructor encouraged me to ask questions.
7. The instructor answered my questions effectively.
8. The instructor was available to provide assistance.

Level 2: Learning

For Level 2, data from instructor-submitted grade book files in Microsoft Excel with test scores, exams, and lab grades are gathered. The number of each type of Level 2 score varies from course to course but is defined in a testing plan document. These data are collected on all "A" School courses. The testing plans, however, do not have any reliability or validity information related to the Level 2 variables. CRESST is not aware if that information exists elsewhere.

Level 3: Behavior

For Level 3, the primary data collected are from "A" School graduates and their supervisors. They are asked to assess the competency, adequacy of training, and task frequency of approximately 30 tasks related to a selected enlisted rating. Competency is defined as the possession of sufficient knowledge and skill, adequacy of training is how well the training in "A" School prepared the Sailor for their job duties, and task frequency is how often the task is performed on the job. The question response options are found in Table 2.

Table 2

Level 3 Response Options for Three Dimensions (Proficiency, Quality of Training, and Frequency)

Dimension	1	2	3	4	5	6	7
Competency	Very Limited	Limited	Somewhat Proficient	Competent	Highly Competent	(not used)	(not used)
Adequacy of Training	Inadequate	Partially Adequate	Adequate	Good	Excellent	(not used)	(not used)
Frequency	Never	Annually	Semi-annually	Quarterly	Monthly	Weekly	Daily

Affective measures are also obtained for Level 3, including one-item measures of interest, self-efficacy, and anxiety (worry) (Lee et al., 2023a, 2023b). Interest is defined as a person's feeling that accompanies or causes special attention to something, in this case, their job (rating); self-efficacy is a person's belief that they can be successful when carrying out a particular task; and anxiety is a person's uncomfortable feeling of nervousness or worry about something that is happening or might happen in the future. The questions, using the Operations Specialist (OS) rating as an example enlisted rating, are as follows:

How interested are you in the OS rating? *(interest item)*

A. Not interested
B. Neutral
C. Interested
D. Very Interested

In general, regarding the OS tasks you perform: *(self-efficacy item)*

A. I could not do them
B. I would struggle to do them on my own

C. I could do them with a bit of effort
D. I could do them easily

In general, I am disappointed with my performance on OS tasks *(anxiety/worry item)*

A. Almost never
B. Sometimes
C. Often
D. Almost always

As these affective measures are one-item scales, reliability cannot be computed. However, validity evidence based on correlational analysis consists of theory-driven predictions and the research literature, where interest and self-efficacy would be positively related to cognitive outcomes, and anxiety (worry) would be negatively related to cognitive outcomes. In addition, anxiety (worry) would be negatively related to each other. Such validity evidence can be found in Lee et al. (2023a, 2023b).

Data Sources

Level 1 and 2 Data

There are usually sufficient data available related to the Level 1 and 2 data sources. For Level 1, by Navy direction, Sailors provide feedback on a course's content and effectiveness of the instructors at the end of a course. These data, however, are anonymous, so it is not possible to correlate them with the Level 2 or Level 3 data. The Level 2 data are the assessments taken throughout the course, and may also include exams, lab work, a culminating activity, and the final grade point average.

Future Level 3 Data

In the future, Level 3 data can be obtained through assessments used in the fleet and shore installations. The Surface Training and Readiness Management System (STRMS; Aptima, 2022) software has already piloted some of this capability for engineering and combat systems watch teams onboard ships, with the Personnel Qualification Standard (PQS) catalog and other on-the-job assessments being stored to monitor the readiness of the watch teams, down to individual Sailors, as assessed by their supervisors onboard. Watch teams are teams of Sailors with a particular mission focus such as combat and watch stations are locations occupied by members of the watch team. STRMS includes the watch team replacement plan (WTRP) tool that keeps track of the watch standers' qualifications and progress towards qualification. Eventually, performance assessments conducted in the fleet could objectively measure Kirkpatrick Level 3 behaviors.

There is ongoing effort in the fleet to incorporate the use of Experience API (xAPI) statements into the Navy systems both onboard and ashore to standardize the format of data to allow for interoperability of the systems and store in process as well as result data. Once the xAPI statements (U.S. Navy, 2023b) are integrated into performance assessments and other exams, this information can be used to track measures, metrics, and indicators that follow the Navy's xAPI profiles. The data sources would be the PQS catalog and time to qualify measures.

According to the PQS catalog, there are over 800 PQSs across the Navy (Naval Education and Training Command, 2021b). The OPNAV Instruction 3500.34H document describes the PQS program:

> The PQS program ensures personnel demonstrate required competencies prior to performing specific duties. PQS delineates the minimum knowledge, skills and abilities that an individual must demonstrate before standing watches or performing other specific duties. These watches or duties will be necessary for the safe, secure and proper operation of a ship, aircraft or support system. PQS is an integral part of the development of the Navy's workforce and the formal process for documenting completion of on-the-job training. (U.S. Navy, 2021, p. 1)

Each PQS includes a description of the contents as follows:

> PQS is divided into three sections. The 100 Section (Fundamentals) contains the fundamental knowledge from technical manuals and other texts necessary to satisfactorily understand the watchstation/workstation duties. The 200 Section (Systems) is designed to acquaint you with the systems you will be required to operate at your watchstation/workstation. The 300 Section (Watchstations) lists the tasks you will be required to satisfactorily perform in order to achieve final PQS qualification for a particular watchstation/workstation. All three sections may not apply to this PQS, but where applicable, detailed explanations are provided at the front of each section. (Naval Education and Training Command, 2013, p. 6)

Note that some PQSs do not have 200 section entries, such as the 3M PQS.

Current Navy practice is that most PQSs are kept in paper format versus electronically. PQS line items and other signatures are graded by supervisors as passed with the inclusion of a signature and date. The final signature for the qualification comes from the Commanding Officer or Designated Representative at the end of the document. Efforts are underway to move to electronic means for capturing this information. For example, the STRMS system includes a place in the WTRP tool for PQS information to be entered. This tool monitors progress toward qualification with graphics to show this information, that is, if the Sailor is on track to become qualified and when that might occur.

Finally, there is a Navy measure referred to as time to qualify (TTQ), which is the amount of time taken to qualify from when a Sailor starts their duties to the date of signoff of the final qualification by the commanding officer (CO) or designated representative, with a service record entry date entered in their electronic service record in the Navy Standard Integrated Personnel System (NSIPS). Estimated completion times are listed under each qualification typically in weeks. According to the NAVEDTRA 43100-1M PQS unit coordinators guide (E. Mejias, personal communication, November 16, 2023):

> Estimated completion times are originally established by the SMEs at the PQS workshop. It is imperative that Department Heads and Division Officers reevaluate specific time periods or time limits for qualification for each qualification requirement. The time period takes into account the ship or squadron's deployment schedule and will tell each individual the time they have to become qualified. The time limit also provides a means whereby the Department Head, Division Officer, and supervisory personnel can check individual progress. This is a goal to instill motivation in each individual and the spirit of competition within the work center. Completion times listed under watchstations are only recommendations and may be adjusted by individual commands.

For example, the estimated time to qualify for Basic Damage Control (306) is four weeks elapsed (Naval Education and Training Command, 2021a). This particular qualification is required for every member of the Navy.

Approaches for Combining Competencies

This section describes the approaches that could be used to combine scores from various measures and metrics to create a composite score. This composite score aims to create a summary score of training effectiveness from multiple metrics to serve as an indicator of quality (i.e., key performance indicator or KPI) for use by management and policymakers (e.g., GAO). This KPI should have the following characteristics:

- Ideally, the indicator would combine the three levels of the Kirkpatrick evaluation model (Kirkpatrick & Kirkpatrick, 2016).
- The indicators may work within individual enlisted ratings and among ratings or clusters of ratings.
- The indicator could be used to compare effects over time and among cohorts.

Current Data Availability

The current NETC data available for analyses can be found in the following table (Table 3).

Table 3
Available Data by Kirkpatrick Level

Kirkpatrick level	Tied to identifiable individuals	Data description
1	No	Course evaluations from graduates related to course content and instructor effectiveness
2	Yes	Course performance measures (multiple knowledge tests, performance tests, and/or lab scores, typically out of 100 points each). A score of 75 or higher is required to pass a given assessment; otherwise, the person must retake the test. If unable to pass again, the person may come under academic review and may be removed from the course or set back (required to retake the course). Only the final score is kept as data.
3	Yes	Feedback on three dimensions related to approximately 30 tasks: competency, adequacy of training, and task frequency. Both individual Sailor's and supervisor's data are available from their questionnaire responses.

In summary, there are usually Level 1 and Level 2 data available on all NETC courses, and there are some Level 3 data on selected enlisted ratings that would permit development of a single composite measure and all competencies to be compared.

Future Data Availability

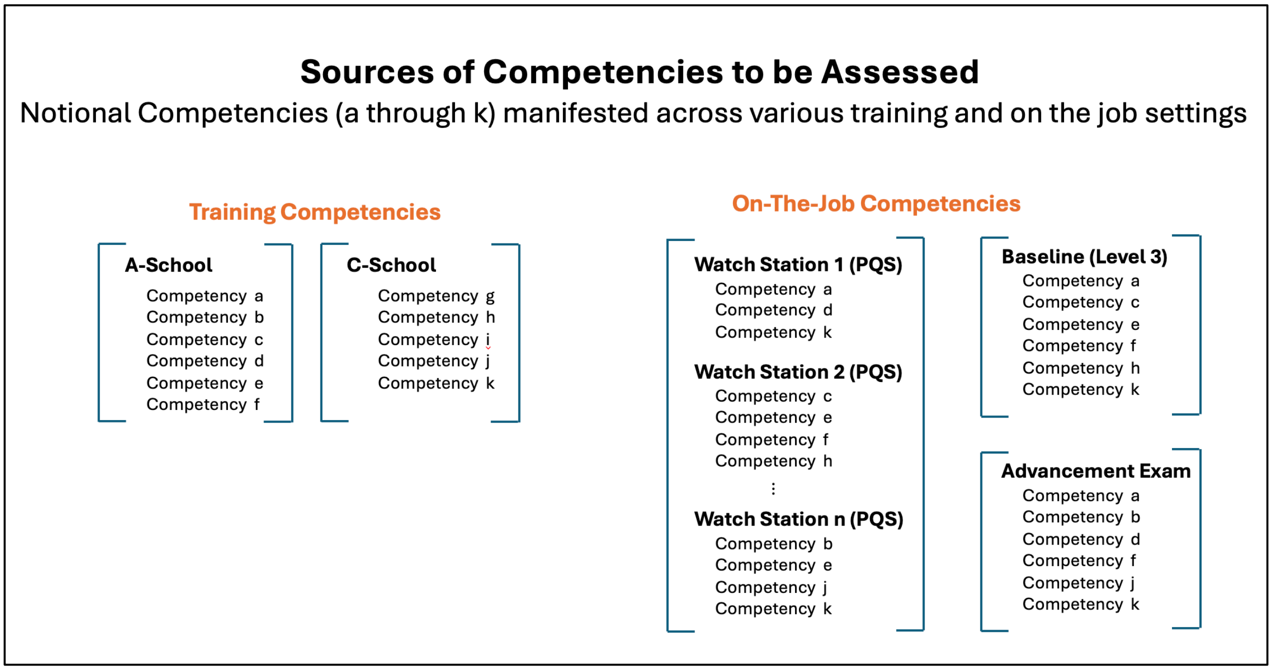
In the future, data such as PQS time to qualify (lower number is better) and in situ assessments like STRMS watch stander scoring can be included as measures in the equation used to create a composite score, with an appropriate weight assigned to each measure depending on the audience and their intended use (purpose) of the composite score or metric.

When considering a metric that can be used to indicate the Navy's overall training effectiveness and readiness, a single common construct is needed that enables valid comparisons across the myriad different units, courses, departments, and mission areas of focus within the Navy. Such a construct is the notion of competencies. The Navy uses the DOD definition of competency as "an observable measurable pattern of knowledge, abilities, skills, and other characteristics that individuals need to successfully perform their work" (U.S. Department of Defense, 2016, p. 21). We can consider competencies as individual, explicitly defined units of skill, knowledge, and attributes that a Sailor can possess at various points throughout their career, which can be measured explicitly through various means.

Figure 1 shows how a notional set of competencies for a given enlisted rating (i.e., Competencies a through k) might be manifested across both training and on-the-job paradigms.

Figure 1

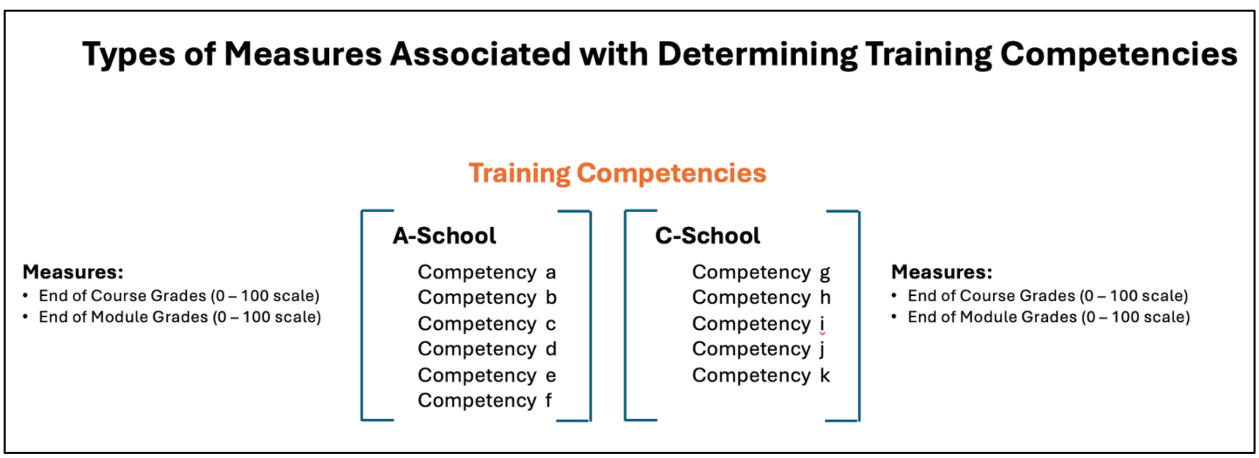
Training and On-the-Job Competencies Mapped to Measures



For training (see Figure 2), competencies can be explicitly mapped to module and course objectives and assessed through measures of a Sailor's end-of-module and end-of-course grades.

Figure 2

Types of Measures Associated With Determining Training Competencies

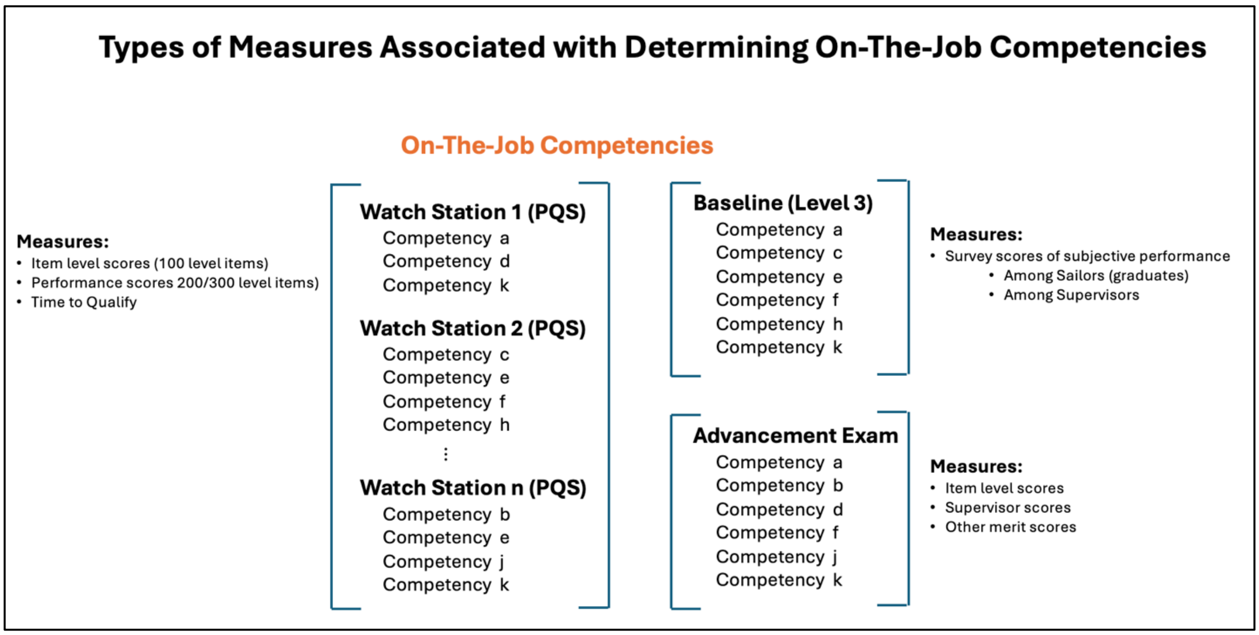


The presumption for training is that what is being trained (expressed as performance objectives) closely aligns with the knowledge, skills, and performance conditions that a Sailor would face when performing such tasks and duties on the job. As such, measures of performance arising from schoolhouse assessments can be considered alongside measures of performance arising from on-the-job activities. Specifically, we can imagine that “A” School is focused on preparing its graduates for tasks and duties that require Competencies a through f. In contrast, “C” School is alternatively focused on Competencies g through k (see Figure 2).

Once on the job (see Figure 3), Sailors can assume various positions, ranging from shore-based to ship-based, depending on where they are assigned (based on the Navy's needs).

Figure 3

Types of Measures Associated With Determining On-the-Job Competencies



Indeed, not every Sailor of a given rating will seek to qualify for the same set of watches (PQS), as ship types (and the equipment used therein) differ across the fleet. As such, we can imagine that the full complement of competencies—represented in our example across “A” School and “C” School—might be distributed among a variety of watchstations, where for any given watchstation, only a subset of these competencies is required.

Qualifying for a given watch requires the Sailor to pass an explicit PQS exam, consisting of 100-level multiple-choice style assessment items and 200- and 300-level performance assessment tasks. As indicated in Figure 3, PQS assessment outcomes associated with each watchstation provide another source of data that can be used to evaluate Sailor competencies.

Similarly, when a Sailor goes up for advancement in their rating, they are required to complete an advancement exam, which consists of a variety of assessment items—each of which can be mapped to explicit competencies (e.g., one or more of Competencies a through k in our example). The item-level scores of these assessment items, coupled with other measures of competencies as evaluated by a Sailor's supervisor, could serve as a further source of data used in our metric.

Finally, the baseline questionnaire work (Lee et al., 2023a, 2023b) that CRESST is conducting is focused on assisting NETC's efforts to obtain fleet feedback on enlisted Sailor task performance (via eliciting questionnaire responses from its "A" School graduates and their supervisors for several Navy ratings). These subjective performance measures among Sailors and their supervisors can also be seen as mapping to explicit competencies and, as such, offer an additional strand of information regarding the state of Sailor competencies within a given rating. As shown in Figure 3, the data currently collected for both the advancement exam and baseline questionnaire data—similar to the watchstations—comprise a subset of the total competencies of the rating (in our example, Competencies a through k). Note: for the baseline project questionnaire data, the information currently collected includes Level 1 anonymous course content and instructor feedback, Level 2 end-of-module (and where available, end-of-course) scores, and Level 3 subjective evaluations of performance among supervisors and graduates. The types and extent of data collected under this baseline effort can be expanded as Navy needs evolve.

The various data sources shown in Figure 1 collectively show the information that can be considered when devising a composite metric that speaks to overall training effectiveness and readiness from the perspective of acquired Sailor competencies. By considering each of these disparate sources of performance, not only is an "apples-to-apples" comparison possible among individual competencies, insights can also be gained as to how those competencies may change over time (e.g., present during training but lost once on the job). Further, considering this information in the aggregate can help highlight potential gaps in training or mismatches regarding training emphases (e.g., spending too much or too little time training a competency compared to its needs on the job).

Weighting Methods

There are several possible weighting methodologies to consider for qualitative and quantitative sources. Qualitative sources pertain to data sources that are expressed in words and used to understand concepts, thoughts, or experiences. Quantitative sources pertain to data represented in numbers and graphs and are used to test or confirm theories and assumptions to establish generalizable facts about a topic (Kyngäs, 2020; Streefkerk, 2023). One constraint in quantitative analysis is that the method or methods chosen will also depend on the number of cases used in the analyses (i.e., they would have sufficient power; Bhandari, 2023). For instance, ideally, a minimum of 1000 cases is needed to use specific methods such as factor analysis. Thus, factor analyses would rarely be done on Navy training effectiveness data.

Qualitative Information

- This scheme relies on expert judgment and consensus about the relative value of the Kirkpatrick levels and item categories or specific items in the responses. Metrics could also be weighted by time or cost of instruction.
- Experts must be given a replicable protocol to ensure they judge according to the same criteria.
  - The reliability and validity of the scores are essential, that is, there is a need for multiple judges, raters, or experts. Interrater agreement (e.g., Krippendorff's alpha) and internal consistency (e.g., Cronbach's alpha) should be reported.
- Proportional weights based on expert judgments and the purpose of the indicator will be assigned based on the type of user.
  - Level 3 might be seen as more critical to the fleet than Level 2, so it would receive a weight reflecting its disparate importance (e.g., a 75% weight) when combined into an indicator. In contrast, NETC might weight Level 2 data as more important to facilitate course revision (and would assign a weight of 25% for Level 3).
  - Should responses of supervisors and individual Sailors be combined, differential weights could be assigned as desired (e.g., supervisor responses count 1.5 times more than those of the individual Sailor).

Quantitative Information

- For the available data at Level 2 or Level 3, the minimum level of cases would be for the following types (with the assumption of medium effect sizes, power = .8, alpha = .05, *SD* of scores = 10, range of scores = 0-100):
  1. Descriptive statistics:
     a. Means (about 20–25 cases would be required)
     b. Correlations (minimum cases would be about 25)
     c. Medians (minimum cases would be about 20–25)
  2. Inferential statistics:
     a. *t* tests (minimum cases required would be 64 per group: 138 total)
     b. ANOVA (minimum cases required would be about 53 per group:158 total)
     c. Regression analysis (minimum number of cases would be about 76)
     d. Almost all multivariate statistics (e.g., structural equation modeling [SEM] or factor analysis) would require a minimum number of more than 1000 cases. It thus would not be used in a vast majority of Navy examples.

- Creating a standard score (e.g., *t* score) **within** an enlisted rating is one matter, but comparing across different ratings may not be feasible. The scale is the same (same *t*-score scale) but the original means of different ratings might differ at the onset. In other words, a standard score is only meaningful within a rating and only across ratings if the standard score is created across a group of similar ratings.
- Sample size issues must be considered. Ideally, 1000 cases are required to create a robust standard score, and similar cases are required to create a robust score based on factor analytic approaches.

Reporting and Communication of Meaning

- Scores can be reported as a standard score, comparing indicators to a mean and including the spread of scores. However, such standard scores would require 1000 cases and thus are not recommended. Instead, if such information is useful across enlisted ratings, we recommend comparing proportions of composite scores above or below some cutoff value.
  - For example, the proportion of trainees that have a composite score greater or less than the mean value can be compared across ratings, and statistical significance can be tested (e.g., using Z-test).
- A brief, common-language description could be appended to each report, describing how the indicator was built and the appropriate allowable comparisons. A summary of this information could be added as an extended footnote to assist the public or less technically adept personnel in interpreting the findings.
- Cut scores could be added to make the data more accessible to interpret as above or below the cut scores.

Reliability and Validity of the Composite Score

The recommended approaches help create a composite metric to measure individual trainees' overall task competencies. Different competencies are currently captured through the assessments. Individual trainees' overall job readiness can be assessed by creating a composite metric to capture multiple competencies. Reliability and validity of the metrics are important considerations when creating composite scores. Reliability relates to the consistency with which the metric captures the intended construct, while validity relates to the extent to which the measure accurately measures the intended construct. Content-oriented validity involves verifying that the metrics adequately capture the appropriate content, established here by verifying the contents of the composite metric with independent experts (American Educational Research Association et al., 2014). In both suggested approaches, multiple experts review the content of the questionnaires and information obtained at each stage of the composite metric development to establish consensus regarding the competencies captured and the weightings for the composite score. Concurrent validity evidence may be defined as the extent of agreement between two conceptually related measures obtained around the same time

(American Educational Research Association et al., 2014; American Psychological Association, n.d.). Concurrent validity evidence is provided by the level of agreement between the two approaches used to develop the composite metric. Composite score weights are determined under both approaches based on information provided by subject-matter experts. The composite metrics are compared to determine the extent of agreement between the two approaches in capturing the intended competencies.

The reliability of the weighted composite score is an important consideration to ensure the developed measures consistently capture the same construct (competency) determined through the various approaches outlined in this proposal. Cronbach's alpha and item-total correlation are examples of measures that quantify internal consistency (a type of reliability). In addition, Kane and Case (2004) developed a reliability measure intended for composite scores. Interrater reliability measures the degree of agreement between raters and experts and is important in mitigating observer bias when eliciting subjective judgments from such raters and experts. Multiple experts are provided the same set of information and training and are independently asked to perform the same tasks, which helps improve interrater reliability. Krippendorff's alpha is also calculated to quantify interrater reliability.

An Example of the Composite Score Using CRESST Data

Quantitative Approach Using Baseline Questionnaires

Multiple questionnaires and rounds of each questionnaire are administered to determine the underlying competency measure(s) and the composite score weighting scheme. The questionnaires aim to develop the structure for the competency measure(s), assign weights to each assessment according to their importance, establish a scoring algorithm for unified measure(s) of competency, and validate the algorithm.

As mentioned above, the baseline questionnaires are administered to the graduates of an "A" School course for a given enlisted rating. The respondents' supervisors also answer similarly worded questions related to their subordinates. The questionnaire measures task performance across three dimensions (or constructs): task frequency, task competency, and adequacy of training provided for each task in "A" School. The number of tasks is approximately 30, determined based on importance and agreed upon by subject-matter experts from the Learning Center, the Type Command (TYCOM), and NETC personnel. These questionnaires are given about six months to a year after graduation per NETC instruction (1540.2C; Naval Education and Training Command, 2023). CRESST is providing the analysis support for this effort (Lee et al., 2023a, 2023b), and it includes a comparison of Kirkpatrick Level 3 questionnaire responses with Kirkpatrick Level 2 outcome data (in-course and end-of-course exams and associated lab or homework assignments).

Example Composite Score Using Kirkpatrick Levels 2 and 3 Data

Decision-makers at higher levels in the Navy may benefit from a composite score related to enlisted Sailor competencies. The purpose of the composite score is to convey information to decision-makers on training effectiveness, including Kirkpatrick Level 2 (learning outcomes) and Level 3 (on-the-job evaluations of behaviors). These scores can be used to compare within and between similar ratings or measure more generalized knowledge, skills, and attributes (KSAs), such as procedural compliance, maintaining safety, or problem-solving between all enlisted ratings. Composite scores can be calculated at the individual student level and be aggregated as needed. In addition, composite score means can provide an overall composite score related to various task competencies.

Representative Report for Various Navy Stakeholders

A representative report is shown in Table 4. Depending on the weighting preferred and approved by Navy subject-matter experts, one of the three composite score columns would be shown. The weighting options include notional examples of (a) 75% Level 2 data, 25% Level 3 data, (b) 50% Level 2 data, 50% Level 3 data, or (c) 25% Level 2 data, 75% Level 3 data. The Navy may decide to use different weightings. The passing score listed could be 75% or another Navy-determined value. Similar tables could be generated for other enlisted ratings, preferably of similar families or clusters. The level of detail below the overall composite score could be provided if helpful, or just overall composite scores reported for groups of similar enlisted ratings. For example, various supply school ratings like Retail Services Specialist (RS), Logistics Specialist (LS), Personnel Specialist (PS), Yeoman (YN), and Utilitiesman (UT) could be categorized together under a Supply category or by the Learning Center, in this case, Center for Service Support (CSS).

The introduction to this report could include the following:

> NETC has established a composite score for enlisted ratings that considers the Kirkpatrick Level 2 (schoolhouse assessment outcomes) and Level 3 (graduate and supervisor evaluations of competency behaviors) data. The table below shows the competency composite score based on a weighting of [(1) 75% Level 2 data, 25% Level 3 data / (2) 50% Level 2 data, 50% Level 3 data / (3) 25% Level 2 data, 75% Level 3 data, depending on which weighting Navy subject-matter experts selected]. The range for the composite score is between 0 and 100, and passing is 75 or higher. Scores under 75 would warrant improvements to training and instruction at "A" School, "C" School (if applicable), or on the job.

Table 4

Retail Services Specialist (RS) Rating Representative Competency Composite Score Report

RS rating task	Competency composite score		
	.75 L2, .25 L3	.50 L2, .50 L3	.25 L2, .75 L3
Overall RS	**68**	**74**	**80**
Identify Retail Services Specialists Actions	66	70	74
Identify Retail Operator References	66	68	72
Identify Laundry Operator References	64	68	72
Provide Customer Service	68	74	80
Perform Stowage Operations	66	70	74
Generate Procurement Forms	66	68	72
Receive Retail Merchandise	64	68	72
Maintain Ship's Stores	68	74	80
Label Prices on Ship's Store Stock Items	66	70	74
Conduct Universal Product Code (UPC) Maintenance	60	58	56
Organize/Visual Merchandise Items	64	66	68
Rotate Store Stock	64	68	72
Perform Breakout Requirements	64	70	74
Secure Group III Spaces and Merchandise	62	64	66
Perform Vending Operations (i.e., Replenish Vending Machines)	64	68	72
Inventory Ship's Store Stock Items	66	70	74
Perform Electronic Point of Sale (EPOS) Operations	66	70	74
Operate Electronic Point of Sale (EPOS) Cash Register	68	74	80
Perform End of Day Procedures	66	72	78
Maintain Shipboard Retail Spaces	64	70	74
Sanitize Retail Spaces	62	66	68
Secure Group IV Spaces	62	64	68
Perform Retail Operator Duties and Retail Operations	64	68	70
Perform Laundry Operator Duties and Laundry Operations	66	72	78

Composite Score Calculation Method

For demonstration purposes, the composite score for each participant (m) across all tasks can be calculated as:

$$(3)\ Combined\ Score_m = \sum_{1}^{n}(w_1 L2_{m,n} + w_2 L3_{m,n})$$

where $\sum_1^i w_i = 1$ and n = number of tasks. w_1 and w_2 correspond to weights on Level 2 and Level 3 scores, respectively. The mapped Level 2 task score ($L2_{m,n}$) is a simple mean of the individual quintile scores for each mapped assessment:

$$(1)\ L2_{m,n} = \frac{\sum_1^j q_{m,j}}{j}$$

where q_j is the quintile score for mapped level 2 assessment j. The Level 3 task score ($L3_{m,n}$) is a weighted mean of supervisor ($s_{m,n}$) and graduate ($g_{m,n}$) ratings:

$$(2)\ L3_{m,n} = \frac{w_1' s_{m,n} + w_2' g_{m,n}}{w_1' + w_2'}$$

For this example, the Level 2 and Level 3 weights are set at (0.75, 0.25), (0.5, 0.5), and (0.25, 0.75). The Navy may use one of the weightings based on whether they want to emphasize Level 2 scores more than Level 3 (75% L2, 25% L3), equally weight them (50% L2, 50% L3), or emphasize Level 3 scores more than Level 2 (25% L2, 75% L3).

The mean Level 2 score is calculated as the simple mean of mapped Level 2 scores for each task (e.g., for the RS rating, Knowledge Test 1 and Final Exam map to Task 1, so the score is an average of the two). Level 2 test scores are grouped into the five Level 3 competency categories for this example. The Level 2 scores are split into quintiles, the lowest quintile of which corresponds to *Very Limited,* and the highest quintile corresponds to *Highly Competent,* so both measures are on a scale from 1 to 5. The mean Level 3 score is a weighted mean of graduate and supervisor ratings for that task. For this example, supervisor responses are weighted 1.5 times higher than graduate responses. The mean Level 2 and Level 3 scores for each task are combined and weighted using the three structures proposed above to produce a composite score for each task, which can be combined into a single composite score for each participant of a particular rating.

For example, to calculate the composite score of 3.0 for Participant RS102 for Task 1: Identify Retail Services Specialists Actions:

3. $L2_{m,n} = \frac{\sum_1^j q_{m,j}}{j} = \frac{2+3}{2} = 2.5$, where Knowledge Test 1 (RS102 scored 85 which is assigned to quintile 2) and Final Exam (RS102 scored 88 which is assigned to quintile 3) are mapped to Task 1 (see Table 5)

4. $L3_{m,n} = \frac{w_1' s_{m,n} + w_2' g_{m,n}}{w_1' + w_2'} = \frac{1.5*4+1*5}{2.5} = 4.4$, where supervisor responses are assigned 1.5 times the weight of graduates. RS102 received 4 from supervisors and 5 from self-rating.

5. $Combined\ Score_m = \sum_1^n (w_1 L2_{m,n} + w_2 L3_{m,n}) = .75 * 2.5 + .25 * 4.4 = 3.0$, where Level 2 scores are weighted 75% and Level 3 scores are weighted 25%.

Table 5

Retail Services Specialist (RS) Rating Level 2 and Level 3 Mapping

Level 3 task	PT	KT1	KT2	KT3	KT4	FE
Identify Retail Services Specialists Actions		X				X
Identify Retail Operator References	X	X				X
Identify Laundry Operator References		X				X
Provide Customer Service	X	X				X
Perform Stowage Operations	X	X				X
Generate Procurement Forms		X				X
Receive Retail Merchandise		X				X
Maintain Ship's Stores	X		X			X
Label Prices on Ship's Store Stock Items	X		X			X
Conduct Universal Product Code (UPC) Maintenance	X		X			X
Organize/Visual Merchandise Items			X			X
Rotate Store Stock			X			X
Perform Breakout Requirements	X		X			X
Secure Group III Spaces and Merchandise	X		X			X
Perform Vending Operations (i.e., Replenish Vending Machines)	X		X			X
Inventory Ship's Store Stock Items			X			X
Perform Electronic Point of Sale (EPOS) Operations	X			X		X
Operate Electronic Point of Sale (EPOS) Cash Register	X			X		X
Perform End of Day Procedures				X		X
Maintain Shipboard Retail Spaces				X		X
Sanitize Retail Spaces				X		X
Secure Group IV Spaces	X				X	X
Perform Retail Operator Duties and Retail Operations			X			X
Perform Laundry Operator Duties and Laundry Operations	X				X	X

Note. PT = Performance Test; KT= Knowledge Test; FE = Final Exam.

Composite Score Tables

The task names referenced in the example composite score tables are listed in Table 6. Table 7 to Table 9 present examples of composite scores for the RS rating using the weighting structures proposed above. Note the weighting structures used are for illustration purposes and may change once more assessments are included and discussions with experts occur. In the tables is a score for each task and participant, which is averaged to form the overall composite score for each participant. The scores range from 1 to 5, with 1 corresponding to *Very Limited* and 5 corresponding to *Highly Competent*. For example, in Table 7, participant RS102 has a competency score of 3.0 on Task 1 and an overall competency score of 3.3 (*Somewhat Competent*) across all 24 tasks. The rightmost column contains overall mean composite scores for each participant, and the bottom row contains overall mean scores for each task. The value in the bottom right cell (e.g., 3.2 in Table 7) represents the class mean across all tasks for the RS rating. Table 10 shows the summary of the prior three tables for comparison purposes. One observation from the tables is that Level 3 scores tend to be higher than Level 2 scores, and thus, a higher weighting on Level 3 scores (e.g., 75% on Level 3 scores and 25% on Level 2 scores) results in higher overall composite scores. The number of participants is low in this example (n = 33) because for the data we had collected, out of 60 initial participants, only 33 had both supervisor and graduate responses.

Comparisons

Comparisons between enlisted ratings would need to be understood with the caveat that ratings will have a different number of variables at Level 2, so a strict comparison is difficult. For variables identified in the schoolhouse as important to the course, this is the mean score on enlisted rating A vs. enlisted rating B, etc. To facilitate direct comparisons of composite scores between ratings, some important considerations must be addressed, such as whether the scores capture the same competencies, are measured in the same manner, and have the same meaning. Further, to group similar ratings for comparison, it is assumed that the samples across ratings represent the same population and that any systematic differences between the samples are attributed to intended variations on the same competencies (May et al., 2009). One could look at the percentage above passing (e.g., 75) for each enlisted rating and compare those percentages. Conversely, the percentage below could also be examined. For example, for the .75 L2, .25 L3 weighting, there were nine out of 33 or 27.2% above 75 and, therefore 72.7% below 75. To compare two enlisted ratings, a two proportion Z-test (Penn State University, 2023) could be used here to look for statistically significant differences.

Decisions Based on Weighting Method Using Level 2 and Level 3 Data Example

Table 11 shows Table 10 as percentages, essentially multiplying the numbers in Table 10 by 20. With respect to decision-making, if 75% is used as a cutoff value (or some other standard setting cutoff value greater than 75% and less than 80%) to assess performance, then overall class mean scores for all tasks using the first two weighting options (68% and 74%) would indicate that the class needs improvement, while the overall mean score on the last weighting (80%) would be considered as “passing” or that training is satisfactory based on the Sailors’ overall competency. In other words, the weighting structure and cutoff value for performance assessment are crucial factors in deciding whether individuals are competent in performing assigned tasks. As Level 3 assessments reflect on-the-job performance, assigning higher weights to Level 3 scores may be more meaningful than Level 2 scores if the focus is on task performance.

Table 6

Retail Services Specialist (RS) Rating Questionnaire Tasks

Task no.	Task name
1	Identify Retail Services Specialists Actions
2	Identify Retail Operator References
3	Identify Laundry Operator References
4	Provide Customer Service
5	Perform Stowage Operations
6	Generate Procurement Forms
7	Receive Retail Merchandise
8	Maintain Ship’s Stores
9	Label Prices on Ship's Store Stock Items
10	Conduct Universal Product Code (UPC) Maintenance
11	Organize/Visual Merchandise Items
12	Rotate Store Stock
13	Perform Breakout Requirements
14	Secure Group III Spaces and Merchandise
15	Perform Vending Operations (i.e., Replenish Vending Machines)
16	Inventory Ship’s Store Stock Items
17	Perform Electronic Point of Sale (EPOS) Operations
18	Operate Electronic Point of Sale (EPOS) Cash Register
19	Perform End of Day Procedures
20	Maintain Shipboard Retail Spaces
21	Sanitize Retail Spaces
22	Secure Group IV Spaces
23	Perform Retail Operator Duties and Retail Operations
24	Perform Laundry Operator Duties and Laundry Operations

Table 7

Composite Competency Scores With 75% Weighting on Level 2 Scores and 25% Weighting on Level 3 Scores

Participant	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	Task Avg
RS102	3.0	3.0	2.7	3.4	3.2	2.7	2.8	3.7	3.7	3.7	3.6	3.7	3.7	3.9	3.7	3.6	2.8	3.0	2.5	2.6	2.5	4.0	3.6	4.0	3.3
RS105	3.6	3.6	3.6	3.7	3.7	3.0	3.7	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	2.6	3.4	2.5	3.3
RS106	3.7	4.0	3.6	4.0	3.9	3.6	3.6	3.0	2.8	2.9	2.5	2.8	3.5	3.5	3.5	2.7	3.6	3.6	3.7	4.3	4.1	4.0	2.8	4.0	3.5
RS107	3.0	2.6	2.7	2.8	2.6	2.4	2.9	2.6	2.8	2.8	2.9	3.0	2.6	2.6	2.6	3.0	2.4	2.4	2.6	2.6	2.8	3.4	3.0	2.7	2.7
RS108	2.4	2.6	2.7	2.7	2.7	2.4	2.7	1.8	1.8	1.7	1.8	2.0	2.1	2.1	2.2	1.9	2.0	1.9	2.3	2.6	2.6	2.0	1.9	1.9	2.2
RS109	2.7	3.0	2.5	3.3	3.1	2.5	2.5	3.1	3.1	3.1	2.6	2.6	3.1	3.1	3.1	2.6	3.6	3.6	3.4	3.4	3.4	3.1	2.6	3.0	3.0
RS111	3.2	3.2	2.7	3.2	2.8	2.7	3.2	3.2	3.2	2.9	3.2	3.1	2.9	2.9	2.9	3.1	2.9	3.2	3.2	3.2	3.2	2.2	3.1	2.0	3.0
RS113	3.4	3.6	3.4	3.6	3.6	3.3	3.4	4.4	4.4	4.4	4.5	4.5	4.4	4.4	4.4	4.5	4.1	4.1	4.1	4.1	4.1	4.4	4.5	4.3	4.1
RS119	2.0	1.9	2.0	2.0	2.0	1.8	1.8	2.0	2.2	1.9	2.1	2.1	2.0	2.3	2.0	2.0	2.4	2.4	2.6	2.9	2.9	2.8	2.0	2.5	2.2
RS120	3.1	2.5	3.0	2.6	2.6	3.0	3.1	2.3	2.2	2.2	2.6	2.7	2.4	2.5	2.3	2.6	2.5	2.5	3.0	3.1	3.1	3.3	2.7	3.2	2.7
RS121	3.8	4.2	3.5	4.3	3.9	3.2	3.0	4.7	4.7	4.0	4.6	4.3	4.6	4.6	4.4	4.6	3.9	4.3	3.7	3.9	3.9	3.7	4.4	3.6	4.1
RS122	4.3	3.5	4.3	3.5	3.5	4.1	4.1	2.8	2.8	2.3	3.1	3.0	2.6	2.6	2.6	3.0	2.9	2.8	3.4	3.4	3.4	2.4	3.0	2.4	3.1
RS125	4.0	4.0	3.7	4.0	4.0	3.6	4.0	3.6	3.2	3.2	3.6	3.6	3.4	3.6	3.6	3.6	3.4	3.4	3.8	3.8	3.8	3.1	3.4	2.7	3.6
RS126	4.8	4.7	4.9	4.9	4.9	4.2	4.6	4.1	4.0	4.0	3.6	3.5	4.0	4.3	4.1	3.7	4.7	4.5	4.7	4.8	5.0	5.0	3.6	5.0	4.4
RS132	3.3	2.8	3.5	2.9	2.8	2.7	3.3	2.8	2.9	2.8	3.4	3.3	2.9	2.7	2.7	3.3	2.9	2.9	3.4	3.4	3.5	2.9	3.4	2.9	3.0
RS137	3.7	3.6	3.7	3.9	3.9	3.6	3.7	4.3	4.1	4.1	4.1	4.1	4.1	4.3	4.1	4.0	3.6	3.6	3.6	3.9	3.9	3.9	4.1	3.7	3.9
RS138	4.7	4.6	4.9	4.6	4.5	4.4	4.6	4.5	4.5	4.1	4.6	4.6	4.6	4.6	4.5	4.6	4.5	4.5	4.4	4.8	4.9	4.6	4.6	4.8	4.6
RS139	1.7	1.7	1.4	2.0	1.4	1.2	1.4	1.8	1.8	1.3	1.6	1.6	1.7	1.8	1.8	1.6	2.3	2.7	2.9	3.2	3.1	2.5	1.4	2.3	1.9
RS144	3.3	3.4	3.4	3.5	3.4	2.9	3.2	3.0	2.8	2.7	3.0	3.0	3.0	3.0	3.0	2.7	2.0	2.1	2.2	2.2	2.2	2.0	3.0	2.1	2.8
RS145	1.8	1.8	1.9	1.9	2.0	1.9	1.9	2.7	2.7	2.3	3.0	3.3	2.9	2.9	2.8	3.4	2.6	2.3	3.0	3.4	3.5	3.3	3.3	3.1	2.6
RS146	3.5	3.3	3.5	3.3	3.3	2.6	3.2	2.8	2.8	2.7	2.8	2.8	2.8	2.8	2.8	2.6	3.3	3.3	3.4	3.5	3.5	2.8	2.8	2.8	3.0
RS149	4.2	4.0	4.1	4.1	4.1	3.7	3.9	3.8	4.0	4.0	3.9	3.8	4.0	4.1	4.0	4.0	3.9	3.7	3.8	4.1	4.0	3.4	4.2	3.1	3.9
RS150	3.7	3.7	3.7	3.9	3.7	3.6	3.4	4.1	4.3	4.1	4.1	4.1	4.1	4.3	4.1	4.1	3.3	3.3	3.7	3.9	3.9	4.3	4.1	4.3	3.9
RS151	2.6	3.0	2.7	3.4	3.1	2.2	2.3	2.3	2.4	2.4	1.5	1.5	2.3	2.6	2.8	1.5	2.4	2.3	1.4	1.4	1.3	3.6	1.5	3.4	2.3

Participant	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	Task Avg
RS156	4.5	3.9	4.3	4.0	3.9	4.0	4.5	3.9	3.9	3.9	4.5	4.5	3.9	3.9	3.9	4.5	4.1	4.0	4.9	4.9	4.9	4.0	4.5	4.0	4.2
RS159	3.9	3.6	3.9	4.0	4.0	3.4	3.9	2.7	2.8	3.0	2.8	2.7	2.9	3.1	2.9	2.8	3.2	3.2	3.3	3.2	3.5	2.5	2.6	2.4	3.2
RS160	2.2	2.7	2.2	2.7	2.7	2.0	2.4	2.6	2.7	2.7	2.4	2.3	2.9	3.0	3.0	2.4	2.7	2.6	2.2	2.4	2.4	2.7	2.4	2.7	2.5
RS162	4.7	4.5	4.6	4.7	4.7	4.0	4.2	4.4	4.7	4.3	4.5	4.7	5.0	4.9	4.9	4.7	4.0	4.0	4.0	4.1	4.5	3.9	4.7	4.1	4.4
RS163	3.9	3.9	3.9	4.2	4.0	4.1	4.2	3.6	3.6	3.6	3.5	3.9	3.6	3.8	3.8	3.6	2.9	2.9	2.7	2.7	3.0	2.5	3.6	2.5	3.5
RS164	2.5	3.1	2.6	3.4	3.2	2.4	2.6	3.8	3.6	3.2	3.0	3.0	3.2	3.5	3.8	3.0	4.0	3.7	3.2	3.4	3.5	4.1	3.0	4.1	3.3
RS166	3.0	3.4	3.0	3.4	3.4	2.7	2.5	3.8	4.0	4.0	3.8	4.0	4.0	4.0	4.0	4.0	3.2	3.2	2.8	2.8	2.8	4.0	4.0	3.8	3.5
RS169	2.8	2.8	2.8	2.8	2.8	2.8	2.8	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1	2.0	2.0	2.0	2.0	2.0	2.0	3.1	2.0	2.7
RS170	2.5	2.5	2.5	2.5	2.5	2.2	2.5	2.8	3.0	2.8	3.4	3.2	3.0	3.1	3.3	3.4	2.6	2.6	2.6	2.6	2.6	2.1	3.4	2.1	2.7
Mean	3.3	3.3	3.2	3.4	3.3	3.0	3.2	3.2	3.2	3.1	3.2	3.3	3.3	3.4	3.3	3.2	3.1	3.1	3.2	3.3	3.4	3.2	3.3	3.1	3.2

Notes. T1 to T24 correspond to Task 1 to Task 24, respectively. Number of participants = 33.

Table 8

Composite Competency Scores With 50% Weighting on Level 2 Scores and 50% Weighting on Level 3 Scores

Participant	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	Task Mean
RS102	3.5	3.0	2.9	3.7	3.4	2.9	3.2	3.7	3.7	3.7	3.7	4.0	3.7	4.0	3.7	3.7	2.9	3.2	2.9	3.2	2.9	3.9	3.7	3.9	3.5
RS105	3.8	3.8	3.8	4.0	4.0	2.5	4.0	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.2	3.7	3.0	3.6
RS106	4.0	4.3	3.7	4.3	4.0	3.7	3.7	2.9	2.6	2.8	2.6	3.1	4.0	4.0	4.0	3.0	3.1	3.1	3.3	4.5	4.2	3.9	3.2	3.9	3.6
RS107	3.5	3.2	2.9	3.5	3.2	2.4	3.3	3.2	3.5	3.5	3.3	3.5	3.2	3.2	3.2	3.5	3.0	3.0	3.2	3.2	3.5	3.7	3.5	2.4	3.2
RS108	2.4	2.7	2.9	2.9	2.9	2.4	2.9	2.2	2.1	2.0	2.2	2.5	2.8	2.8	3.0	2.4	1.9	1.7	2.6	3.2	3.2	2.6	2.4	2.4	2.5
RS109	3.3	3.3	3.0	3.8	3.5	3.0	3.0	3.5	3.5	3.5	3.2	3.2	3.5	3.5	3.5	3.2	3.9	3.9	3.7	3.7	3.7	3.5	3.2	3.3	3.4
RS111	3.4	3.4	2.3	3.4	2.6	2.3	3.4	3.4	3.4	2.8	3.4	3.2	2.8	2.8	2.7	3.2	2.8	3.4	3.4	3.4	3.4	2.3	3.2	2.0	3.0
RS113	3.7	3.9	3.7	3.9	3.9	3.5	3.7	4.4	4.4	4.4	4.5	4.5	4.4	4.4	4.4	4.5	4.2	4.2	4.2	4.2	4.2	4.4	4.5	4.2	4.2
RS119	2.5	2.4	2.5	2.7	2.7	2.2	2.2	2.7	3.0	2.4	2.8	2.8	2.7	3.2	2.7	2.5	2.7	2.7	2.7	3.3	3.3	3.5	2.5	3.0	2.7
RS120	3.1	2.6	2.9	2.8	2.8	2.9	3.1	2.6	2.4	2.4	2.7	2.9	2.7	2.9	2.6	2.7	2.6	2.6	2.9	3.2	3.1	3.3	2.9	3.1	2.8
RS121	4.1	4.3	3.6	4.5	3.8	3.0	2.5	4.6	4.6	3.3	4.8	4.1	4.5	4.4	4.1	4.8	3.7	4.5	3.9	4.3	4.3	3.6	4.3	3.4	4.0
RS122	4.5	4.0	4.5	4.0	4.0	4.2	4.2	3.5	3.5	2.6	3.8	3.5	3.2	3.2	3.2	3.5	3.4	3.2	3.7	3.7	3.7	3.0	3.5	3.0	3.6
RS125	3.9	3.9	3.3	3.9	3.9	3.1	3.9	3.7	2.9	2.9	3.7	3.7	3.3	3.7	3.7	3.7	2.8	2.8	3.6	3.6	3.6	3.1	3.4	2.3	3.4
RS126	4.6	4.3	4.7	4.7	4.7	3.3	4.1	4.2	4.0	4.0	3.8	3.6	4.0	4.5	4.2	4.0	4.4	4.0	4.3	4.5	5.0	5.0	3.8	5.0	4.3
RS132	3.5	3.2	4.0	3.5	3.2	2.4	3.5	3.2	3.5	3.2	3.7	3.5	3.5	3.1	3.0	3.5	3.5	3.5	3.8	3.8	4.0	3.5	3.8	3.5	3.4
RS137	4.0	3.8	4.0	4.3	4.3	3.8	4.0	4.5	4.2	4.2	4.2	4.2	4.2	4.5	4.2	3.9	3.7	3.7	3.7	4.3	4.3	4.3	4.2	4.0	4.1
RS138	4.4	4.5	4.7	4.5	4.2	3.7	4.1	4.2	4.2	3.4	4.1	4.1	4.5	4.5	4.2	4.1	4.2	4.2	3.8	4.5	4.7	4.5	4.2	4.8	4.3
RS139	2.4	2.0	1.7	2.6	1.4	1.3	1.8	2.3	2.3	1.2	2.1	2.1	2.0	2.3	2.3	2.1	1.8	2.6	2.8	3.4	3.1	2.7	1.8	2.2	2.2
RS144	3.5	3.7	3.7	4.0	3.7	2.8	3.4	3.5	3.1	3.0	3.5	3.5	3.5	3.5	3.5	3.0	2.6	2.8	3.0	3.0	3.0	2.5	3.5	2.7	3.2
RS145	2.2	2.2	2.4	2.4	2.6	2.3	2.3	3.0	3.0	2.3	3.0	3.6	3.4	3.4	3.2	3.7	2.8	2.3	3.0	3.8	4.0	3.8	3.6	3.5	3.0
RS146	4.0	3.8	4.0	3.8	3.8	2.2	3.4	3.5	3.5	3.3	3.5	3.5	3.5	3.5	3.5	3.2	3.8	3.8	3.8	4.0	4.0	3.5	3.5	3.5	3.6
RS149	4.3	4.0	4.2	4.2	4.2	3.4	3.7	3.6	3.9	3.9	3.7	3.5	3.9	4.2	3.9	3.9	3.7	3.3	3.5	4.2	3.9	3.7	4.3	3.1	3.8
RS150	4.0	4.0	4.0	4.3	4.0	3.7	3.4	4.2	4.5	4.2	4.2	4.2	4.2	4.5	4.2	4.2	3.1	3.1	4.0	4.3	4.3	4.5	4.2	4.5	4.1
RS151	2.7	2.7	2.9	3.4	2.9	2.0	2.2	1.8	2.0	2.0	1.6	1.6	1.8	2.5	2.9	1.5	2.5	2.3	1.8	1.8	1.6	3.9	1.6	3.4	2.3
RS156	4.5	4.0	4.2	4.3	4.0	3.6	4.5	4.0	4.0	4.0	4.5	4.5	4.0	4.0	4.0	4.5	4.2	3.9	4.7	4.7	4.7	3.9	4.5	3.9	4.2

Participant	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	Task Mean
RS159	3.7	3.2	3.7	4.0	4.0	2.7	3.7	3.0	3.1	3.5	3.2	3.0	3.3	3.8	3.3	3.2	3.3	3.3	3.5	3.4	4.0	3.0	2.8	2.8	3.3
RS160	2.4	2.7	2.4	2.7	2.7	2.0	2.7	2.5	2.7	2.7	2.7	2.5	3.0	3.2	3.2	2.7	2.7	2.5	2.4	2.7	2.7	2.7	2.7	2.7	2.7
RS162	4.4	4.0	4.1	4.4	4.4	3.0	3.4	3.7	4.3	3.6	3.9	4.4	5.0	4.7	4.7	4.4	3.5	3.5	3.5	3.8	4.5	3.8	4.3	4.2	4.1
RS163	3.8	3.8	3.7	4.3	4.0	4.2	4.3	3.8	3.8	3.8	3.5	4.3	3.8	4.1	4.1	3.8	3.3	3.3	3.0	3.0	3.6	3.0	3.8	3.0	3.7
RS164	2.9	3.1	3.2	3.7	3.4	2.7	3.2	4.2	3.9	3.1	3.5	3.5	3.0	3.7	4.2	3.5	4.3	3.6	3.3	3.7	4.0	4.2	3.5	4.2	3.6
RS166	3.5	3.7	3.5	3.7	3.7	2.9	2.6	3.6	3.9	3.9	3.6	3.9	3.9	3.9	3.9	3.9	3.4	3.4	3.2	3.2	3.2	3.9	3.9	3.6	3.6
RS169	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.0	3.0	3.0	3.0	3.0	3.0	3.8	3.0	3.5
RS170	2.9	2.9	2.9	2.9	2.9	2.3	2.9	2.9	3.2	2.9	3.7	3.4	3.2	3.5	3.8	3.7	3.2	3.2	3.2	3.2	3.2	2.9	3.7	2.9	3.2
Mean	3.5	3.4	3.4	3.7	3.5	2.9	3.3	3.4	3.5	3.2	3.4	3.5	3.5	3.7	3.6	3.5	3.3	3.2	3.4	3.6	3.7	3.5	3.5	3.3	3.4

Notes. T1 to T24 correspond to Task 1 to Task 24, respectively. Number of participants = 33.

Table 9

Composite Competency Scores With 25% Weighting on Level 2 Scores and 75% Weighting on Level 3 Scores

Participant	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	Task Mean
RS102	3.9	3.0	3.0	4.1	3.6	3.0	3.5	3.8	3.8	3.8	3.7	4.2	3.8	4.2	3.8	3.7	3.1	3.5	3.4	3.8	3.4	3.9	3.7	3.9	3.6
RS105	3.9	3.9	3.9	4.2	4.2	1.9	4.2	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	3.8	4.1	3.5	3.9
RS106	4.2	4.7	3.7	4.7	4.2	3.7	3.7	2.9	2.4	2.7	2.6	3.3	4.5	4.5	4.5	3.2	2.7	2.7	3.0	4.8	4.3	3.9	3.5	3.9	3.7
RS107	3.9	3.8	3.0	4.3	3.8	2.3	3.6	3.8	4.3	4.3	3.6	3.9	3.8	3.8	3.8	3.9	3.7	3.7	3.8	3.8	4.3	4.1	3.9	2.1	3.7
RS108	2.3	2.7	3.0	3.0	3.0	2.3	3.0	2.5	2.3	2.2	2.5	2.9	3.4	3.4	3.7	2.8	1.9	1.6	2.9	3.8	3.8	3.1	2.8	2.8	2.8
RS109	4.0	3.7	3.5	4.4	4.0	3.5	3.5	4.0	4.0	4.0	3.8	3.8	4.0	4.0	4.0	3.8	4.1	4.1	4.1	4.1	4.1	4.0	3.8	3.7	3.9
RS111	3.6	3.6	2.0	3.6	2.4	2.0	3.6	3.6	3.6	2.7	3.6	3.3	2.7	2.7	2.6	3.3	2.7	3.6	3.6	3.6	3.6	2.5	3.3	2.0	3.1
RS113	4.1	4.1	4.1	4.1	4.1	3.8	4.1	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.3	4.3	4.3	4.3	4.3	4.4	4.4	4.1	4.3
RS119	2.9	2.9	2.9	3.3	3.3	2.5	2.5	3.3	3.8	2.9	3.4	3.4	3.3	4.1	3.3	2.9	3.1	3.1	2.7	3.6	3.6	4.3	2.9	3.5	3.2
RS120	3.2	2.7	2.9	3.0	3.0	2.9	3.2	2.9	2.6	2.6	2.7	3.0	3.1	3.4	2.9	2.7	2.7	2.7	2.9	3.3	3.2	3.2	3.0	2.9	2.9
RS121	4.3	4.5	3.6	4.8	3.7	2.7	1.9	4.6	4.6	2.7	4.9	3.8	4.5	4.3	3.9	4.9	3.6	4.8	4.0	4.6	4.6	3.6	4.1	3.3	4.0
RS122	4.8	4.5	4.8	4.5	4.5	4.3	4.3	4.3	4.3	2.9	4.4	3.9	3.8	3.8	3.8	3.9	3.9	3.6	4.1	4.1	4.1	3.7	3.9	3.7	4.1
RS125	3.9	3.9	3.0	3.9	3.9	2.7	3.9	3.7	2.5	2.5	3.7	3.7	3.1	3.7	3.7	3.7	2.2	2.2	3.4	3.4	3.4	3.2	3.3	2.0	3.3
RS126	4.4	4.0	4.6	4.6	4.6	2.5	3.7	4.3	4.0	4.0	3.9	3.6	4.0	4.8	4.3	4.2	4.1	3.5	4.0	4.3	5.0	5.0	3.9	5.0	4.2
RS132	3.8	3.6	4.5	4.0	3.6	2.1	3.8	3.6	4.0	3.6	4.1	3.8	4.0	3.4	3.3	3.8	4.0	4.0	4.2	4.2	4.5	4.0	4.2	4.0	3.8
RS137	4.2	3.9	4.2	4.6	4.6	3.9	4.2	4.8	4.3	4.3	4.3	4.3	4.3	4.8	4.3	3.9	3.7	3.7	3.7	4.6	4.6	4.6	4.3	4.2	4.3
RS138	4.1	4.5	4.6	4.5	4.0	3.1	3.7	4.0	4.0	2.8	3.7	3.7	4.5	4.5	4.0	3.7	4.0	4.0	3.2	4.3	4.6	4.5	3.8	4.9	4.0
RS139	3.1	2.3	2.1	3.2	1.4	1.5	2.2	2.7	2.7	1.1	2.7	2.7	2.3	2.7	2.7	2.7	1.4	2.6	2.7	3.6	3.2	2.8	2.2	2.1	2.4
RS144	3.8	4.1	4.1	4.5	4.1	2.7	3.6	3.9	3.3	3.2	3.9	3.9	3.9	3.9	3.9	3.2	3.1	3.4	3.7	3.7	3.7	2.9	3.9	3.2	3.6
RS145	2.5	2.6	2.8	2.9	3.2	2.6	2.6	3.3	3.3	2.2	3.0	3.9	3.9	3.9	3.6	4.1	3.0	2.2	3.0	4.2	4.5	4.4	3.9	4.0	3.3
RS146	4.5	4.4	4.5	4.4	4.4	1.8	3.6	4.3	4.3	4.0	4.3	4.3	4.3	4.3	4.3	3.8	4.4	4.4	4.2	4.5	4.5	4.3	4.3	4.3	4.2
RS149	4.5	4.0	4.3	4.3	4.3	3.1	3.6	3.4	3.9	3.9	3.6	3.3	3.9	4.3	3.9	3.9	3.6	3.0	3.3	4.3	3.9	4.1	4.5	3.2	3.8
RS150	4.2	4.2	4.2	4.6	4.2	3.7	3.3	4.3	4.8	4.3	4.3	4.3	4.3	4.8	4.3	4.3	2.8	2.8	4.2	4.6	4.6	4.8	4.3	4.8	4.2
RS151	2.7	2.3	3.0	3.4	2.6	1.7	2.0	1.4	1.7	1.7	1.6	1.6	1.4	2.5	3.1	1.4	2.5	2.2	2.2	2.2	1.9	4.1	1.6	3.4	2.3
RS156	4.4	4.2	4.0	4.7	4.2	3.1	4.4	4.2	4.2	4.2	4.4	4.4	4.2	4.2	4.2	4.4	4.3	3.9	4.6	4.6	4.6	3.9	4.4	3.9	4.2

Participant	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	Task Mean
RS159	3.6	2.8	3.6	4.0	4.0	2.1	3.6	3.2	3.3	3.9	3.5	3.2	3.6	4.4	3.6	3.5	3.5	3.5	3.8	3.6	4.5	3.5	2.9	3.2	3.5
RS160	2.6	2.8	2.6	2.8	2.8	2.0	3.1	2.5	2.8	2.8	3.1	2.8	3.2	3.5	3.5	3.1	2.8	2.5	2.6	3.1	3.1	2.8	3.1	2.8	2.8
RS162	4.1	3.5	3.7	4.1	4.1	2.0	2.6	3.1	4.0	2.9	3.4	4.1	5.0	4.6	4.6	4.1	2.9	2.9	2.9	3.4	4.4	3.7	4.0	4.3	3.7
RS163	3.7	3.7	3.6	4.5	4.0	4.3	4.5	3.9	3.9	3.9	3.4	4.6	3.9	4.3	4.3	3.9	3.6	3.6	3.2	3.2	4.1	3.5	3.9	3.5	3.9
RS164	3.4	3.2	3.8	4.1	3.6	3.1	3.8	4.6	4.1	2.9	3.9	3.9	2.8	3.8	4.6	3.9	4.7	3.6	3.5	4.1	4.5	4.3	3.9	4.3	3.8
RS166	3.9	4.1	3.9	4.1	4.1	3.0	2.6	3.4	3.9	3.9	3.4	3.9	3.9	3.9	3.9	3.9	3.6	3.6	3.5	3.5	3.5	3.9	3.9	3.4	3.7
RS169	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.0	4.0	4.0	4.0	4.0	4.0	4.4	4.0	4.2
RS170	3.4	3.4	3.4	3.4	3.4	2.5	3.4	3.1	3.5	3.1	4.1	3.6	3.5	4.0	4.4	4.1	3.8	3.8	3.8	3.8	3.8	3.6	4.1	3.6	3.6
Mean	3.7	3.6	3.6	4.0	3.7	2.8	3.4	3.6	3.7	3.3	3.6	3.7	3.7	4.0	3.9	3.7	3.4	3.4	3.5	3.9	4.0	3.8	3.7	3.5	3.6

Notes. T1 to T24 correspond to Task 1 to Task 24, respectively. Number of participants = 33.

Table 10

Composite Competency Score Means With Various Weighting Options

Weighting (L2, L3)	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	Task Mean
.75, .25	3.3	3.3	3.2	3.4	3.3	3.0	3.2	3.2	3.2	3.1	3.2	3.3	3.3	3.4	3.3	3.2	3.1	3.1	3.2	3.3	3.4	3.2	3.3	3.1	3.2
.50, .50	3.5	3.4	3.4	3.7	3.5	2.9	3.3	3.4	3.5	3.2	3.4	3.5	3.5	3.7	3.6	3.5	3.3	3.2	3.4	3.6	3.7	3.5	3.5	3.3	3.4
.25, .75	3.7	3.6	3.6	4.0	3.7	2.8	3.4	3.6	3.7	3.3	3.6	3.7	3.7	4.0	3.9	3.7	3.4	3.4	3.5	3.9	4.0	3.8	3.7	3.5	3.6

Notes. T1 to T24 correspond to Task 1 to Task 24, respectively. Number of participants = 33.

Table 11

Composite Competency Score Means With Various Weighting Options as Percentages

Weighting (L2, L3)	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	Task Mean
.75, .25	66	66	64	68	66	66	64	68	66	60	64	64	64	62	64	66	66	68	66	64	62	62	64	66	68
.50, .50	70	68	68	74	70	68	68	74	70	58	66	68	70	64	68	70	70	74	72	70	66	64	68	72	74
.25, .75	74	72	72	80	74	72	72	80	74	56	68	72	74	66	72	74	74	80	78	74	68	68	70	78	80

Notes. T1 to T24 correspond to Task 1 to Task 24, respectively. Number of participants = 33.

Summary

The report discussed developing a composite metric from various U.S. Navy training programs to allow for effectiveness comparisons. It outlined using three levels of the Kirkpatrick training evaluation model currently implemented by the Navy. Level 1 evaluates reactions through anonymous course evaluations. Level 2 assesses learning via test/exam scores. Level 3 examines behavior through questionnaires given to graduates and supervisors regarding task performance. The goal was to create a composite score incorporating these three levels to serve as a key performance indicator for workforce management. This single metric would consolidate data collected from different sources on training outcomes.

Several proposed methodologies for calculating a composite score were described. Both qualitative and quantitative methods could be employed. Various potential Navy data sources were identified that may feed into the composite score calculation. These include PQS time to qualify, STRMS assessments, advancement exam results, and Kirkpatrick Level 3 baseline questionnaires administered by the Navy. Reliability and validity are essential considerations for the composite metric. The expert weighting of the individual measures may help demonstrate the score captures competencies as intended. A worked example using CRESST baseline questionnaire data was provided to test the methodologies. Overall, this report outlined developing a valid and reliable composite score incorporating different competency and training outcome measures into a single metric for evaluation purposes across the Navy training continuum.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

American Psychological Association. (n.d.). Concurrent validity. In *APA dictionary of psychology.* Retrieved January 17, 2024, from https://dictionary.apa.org/concurrent-validity

Aptima. (2022, June 8). *Aptima announces latest phase in Surface Training and Readiness Management System for U.S. Navy*. Retrieved December 19, 2023, from https://www.aptima.com/aptima-announces-latest-phase-in-surface-training-and-readiness-management-system-for-u-s-navy/

Bhandari, P. (2023, June 22). *Statistical power and why it matters: A simple introduction.* Retrieved January 15, 2024, from https://www.scribbr.com/statistics/statistical-power/

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*(3), 221-240.

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation*. Association for Talent Development.

Kyngäs, H. (2020). Qualitative research and content analysis. In H. Kyngäs, K. Mikkonen, & M. Kääriäinen (Eds.), *The application of content analysis in nursing science research* (pp. 3-11). Springer. https://doi.org/10.1007/978-3-030-30199-6_1

Lee, J. J., Senteney, M., Griffin, N., Baker, E. L., & O'Neil, H. F. (2023a). *Analysis and reporting of Level 3 questionnaires and associated Level 1 and 2 data (Phase 1): Option Year 1* (Deliverable Item No. 6 to funder). University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Lee, J. J., Senteney, M., Griffin, N., Baker, E. L., & O'Neil, H. F. (2023b). *Analysis and reporting of Level 3 questionnaires and associated Level 1 and 2 data (Phase 2): Option Year 1* (Deliverable Item No. 7 to funder). University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009-013). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Naval Education and Training Command. (2013, January). *Personnel Qualification Standard for 3-M* (NAVEDTRA 43241-K).

Naval Education and Training Command. (2021a, October 4). *Personnel Qualification Standard for Damage Control (DC)* (NAVEDTRA 43119-N).

Naval Education and Training Command. (2021b, October 8). *Personnel Qualification Standard: PQS catalog*.

Naval Education and Training Command. (2023, March 22). *Naval Education and Training Command Training Effectiveness Program.* (NETCINST 1540.2C)

Penn State University. (2023). *Comparing two proportions.* Retrieved from https://online.stat.psu.edu/stat415/lesson/9/9.4

Streefkerk, R. (2023, June 22). *Qualitative vs. quantitative research: Differences, examples, and methods.* Retrieved January 4, 2024, from https://www.scribbr.com/methodology/qualitative-quantitative-research/

U.S. Department of Defense. (2016, June). *DoD civilian personnel management system: Civilian strategic human capital planning (SHCP)* (DoDI Number 1400.25, Volume 250). http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/140025/140025_vol250.pdf

U.S. Navy. (2021). *OPNAV Instruction 3500.34H. Personnel Qualification Standards program.* Office of the Chief of Naval Operations.

U.S. Navy. (2023a). *NAVINST1540-2C: Naval Education and Training Command Training Effectiveness Program.* N7.

U.S. Navy. (2023b). *Navy performance assessment profile: Requirements and guidelines for implementing xAPI in performance assessments.* Document Version 1.0. Retrieved October 27, 2023, from https://netc.usalearning.net/xapi-library/assessment.html

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)

School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
SE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522

(310) 206-1532
www.cresst.org