Evolution of the SCORE Framework to Enhance Field-Based Performance Evaluations of Emerging Technologies

Brian A. Weiss and Craig Schlenoff
National Institute of Standards and Technology
100 Bureau Drive, MS 8230
Gaithersburg, Maryland 20899 USA
+1.301.975.4373, +1.301.975.3456

brian.weiss@nist.gov, craig.schlenoff@nist.gov

ABSTRACT

NIST has developed the System, Component, and Operationally-Relevant Evaluations (SCORE) framework as a formal guide for designing evaluations of emerging technologies. SCORE captures both technical performance and end-user utility assessments of systems and their components within controlled and realistic environments. Its purpose is to present an extensive (but not necessarily exhaustive) picture of how a system would behave in a realistic operating environment. The framework has been applied to numerous evaluation efforts over the past three years producing valuable quantitative and qualitative metrics. This paper will present the building blocks of the SCORE methodology including the system goals and design criteria that drive the evaluation design process. An evolution of the SCORE framework in capturing utility assessments at the capability level of a system will also be presented. Examples will be shown of SCORE's successful application to the evaluation of the soldier-worn sensor systems and two-way, free-form spoken language translation technologies.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: *measurement techniques, modeling techniques, performance attributes.*

General Terms

Measurement, Documentation, Performance, Experimentation, Verification.

Keywords

SCORE, DARPA, ASSIST, TRANSTAC, performance evaluation, elemental tests, vignette tests, task tests, speech translation, soldier-worn sensor.

1. INTRODUCTION

As intelligent systems emerge and take shape, it is important to understand their capabilities and limitations. Evaluations are a means to assess both quantitative technical performance and qualitative end-user utility. System, Component and Operationally Relevant Evaluations (SCORE) is a unified set of criteria and software tools for defining a performance evaluation approach for intelligent systems. It provides a comprehensive evaluation blueprint that assesses the technical performance of a system and its components through isolating and changing variables as well as capturing end-user utility of the system in realistic use-case environments. SCORE is unique in that:

- It is applicable to a wide range of technologies, from manufacturing to defense systems
- Elements of SCORE can be decoupled and customized based upon evaluation goals
- It has the ability to evaluate a technology at various stages of development, from conceptual to full maturation
- It combines the results of targeted evaluations to produce an extensive picture of a systems' capabilities and utility

Section 2 introduces the SCORE framework and its initial evaluation design structure. Section 3 presents SCORE's first applications in evaluating technologies developed under the Defense Advanced Research Projects Agency's (DARPA) Advanced Soldier Sensor Information System and Technology (ASSIST), Phase I and II program along with DARPA's Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) Phase II program. Section 4 discusses the evolution of the framework necessitated by the advancing goals of the ASSIST and TRANSTAC programs. Section 5 describes some future efforts (outside of the above military-based programs) that are expected to use the SCORE framework. Section 6 concludes the paper.

2. BACKGROUND

2.1 SCORE Development

Intelligent systems tend to be complex and non-deterministic, involving numerous components that are jointly working together to accomplish an overall goal. Existing approaches to measuring such systems often focus on evaluating the system as a whole or individually evaluating some of the components under very controlled, but limited, conditions. These approaches do not comprehensively and quantitatively assess the impact of variables such as environmental variables (e.g, weather) and system variables (e.g., processing power, memory size) on the system's overall performance. The SCORE framework, with its comprehensive evaluation criteria and software tools, is developed to enhance the ability to quantitatively and qualitatively evaluate intelligent systems at the component level -- and the system level -- in both controlled and operationally-relevant environments.

SCORE leverages the multi-level Steves/Scholtz evaluation framework that defines metrics and measures in the context of

PerMIS'08, August 19–21, 2008, Gaithersburg, MD, USA.
ACM ISBN 978-1-60558-293-1/08/08.

system goals and evaluation objectives, and combines these assessments for an overall evaluation of a system [1]. SCORE takes the framework a step further by identifying specific system goals and areas of interest. It is built around the premise that, in order to get a comprehensive picture of how a system performs in its actual use-case environment, technical performance should be evaluated at the component and system levels [2]. Additionally, system level utility assessments should be performed to gain an understanding of the value the system provides to the end-users. SCORE defines three evaluation goal types:

- *Component Level Testing – Technical Performance* – This evaluation type involves decomposing a system into components to isolate those subsystems that are critical to system operation. Ideally, all of the components together, should include all facets of the system and yield a complete evaluation.
- *System Level Testing – Technical Performance* – This evaluation type is intended to assess the system as a whole, but in an ideal environment where test variables can be isolated and controlled. The benefit is that tests can be performed using a combination of test variables and parameters, where relationships can be determined between system behavior and these variables and parameters based upon the technical performance analysis.
- *System Level Testing – Utility Assessments* – This evaluation class assesses a system's utility, where utility is defined as the value the application provides to the end-user. In addition, usability is assessed which includes effectiveness, learnability, flexibility, and user attitude towards the system. The advantage of this evaluation mode is that system's utility and value can still be addressed even when the system design and user-interface are not yet finalized (i.e. the working version in place is not perfected).

For each of these three goal types, the following evaluation elements are pertinent:

- Identification of the system or component to be assessed
- Definition of the goal/objective(s)/metrics/ measures
    - Goal – For a particular assessment, the goal is influenced by whether the intent of the evaluation is to inform or validate the system design. The state of system maturity also weighs heavily on the goal specification
    - Objectives – Evaluation objectives are used to separate evaluation concerns. These evaluation concerns also include identifying how different variables impact system performance and determining which should be fixed and which should be modified during testing.
    - Metrics/measures – Depending upon the type of evaluation, either technical performance metrics or utility metrics would be employed.
- Specification of the testing environment(s) – Selecting a testing environment is influenced by a range of aspects including system maturity, intended use-case environments, physical issues, site suitability, etc.
- Identification of participants – The system users, whether they are the technology developers and/or end-users needs to be determined. Actors that will be indirectly interacting with the system through role-playing within the environment also need to be identified.
- Specification of participant training – Technology users must be properly instructed (and have time to practice) on how to appropriately interact/engage the systems. Likewise, the environmental actors require guidance as to how they should perform throughout the test(s).
- Specification of data collection methods – As measures and metrics are specified, data capture methods must be formulated.
- Specification of the use-case scenarios – The evaluation architect must devise the use scenario(s) under which the system (or component) will be tested.

Considering each of these evaluation elements, SCORE takes a tiered approach to measuring the performance of intelligent systems. At the lowest level, SCORE uses component level tests to isolate specific components and then systematically modifies variables that could affect the performance of that component to determine those variables' impact. Typically, this is performed for each relevant component within the system. At the next level, the overall system is tested in a highly structured environment to understand the performance of individual variables on the system. Lastly, the technology is immersed in a richer scenario that evokes typical situations and surroundings in which the end-user is asked to perform an overall mission or procedure in a highly-relevant environment which stresses the overall system's capabilities. Formal surveys and semi-structured interviews are used to assess the usefulness of the technology to the end-user.

3. INITIAL APPLICATIONS

SCORE was initially applied to intelligent systems developed under the DARPA ASSIST and TRANSTAC programs. The SCORE-based evaluations also provided the researchers and end-users with the information needed to determine if and when the technology will be ready for actual use. The SCORE framework identified various key components of the system and evaluated them both independently and as a whole, thus helping to determine the impact of the individual components on the performance of the overall system. This detailed analysis allowed the evaluation team, and the sponsor, to more accurately target the aspects of the systems that were shown to provide the greatest benefit to the overall advancement of the technology. Prior to adopting SCORE, DARPA did not have this level of necessary detail about system and component performance.

3.1 ASSIST – Phase I and II

The DARPA ASSIST program is an advanced technology research and development program whose objective is to exploit soldier-worn sensors to augment a Soldier's mission recall and reporting capability to enhance situational knowledge within Military Operations in Urban Terrain (MOUT) environments [3]. This program is split into two tasks with the NIST Independent Evaluation Team (IET) focused on evaluating task 2 technology. This task stresses passive collection and automated activity/object recognition capabilities in the form of algorithms, software, and tools that will undergo system integration in future efforts.

The process of applying the SCORE framework to the ASSIST evaluations begins with identifying the specific technologies. The technologies were developed by three different research teams. It should be noted that there is no single, fully-integrated ASSIST system, so each team focused their attention on some unique and/or overlapping technologies. The Phase I and Phase II capabilities are broken out as follows:

- Image/Video Data Analysis Capabilities
    - Object Detection/ Image Classification (Phase I)
    - Arabic Text Translation (Phase I)
    - Face Recognition and Matching (Phase II)
- Audio Data Analysis Capabilities
    - Sound Recognition/Speech Recognition (Phase I)
    - Shot Localization/Weapon Classification (Phase I)
- Soldier Activity Data Analysis Capabilities
    - Soldier State Identification/Localization (Phase I and II)

Further explanation of these technologies can be found in [3] and [4]. The next crucial step is to determine the evaluation goals/ objectives and metrics/measures. As outlined by DARPA, at a high level, they are:

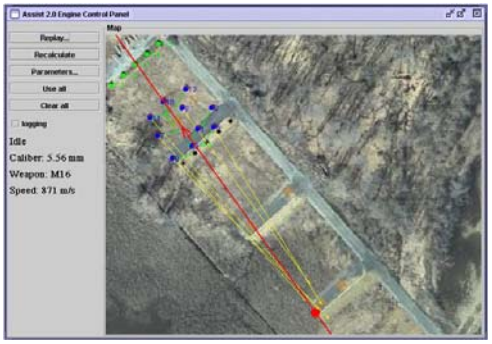
1. The accuracy of object/event/activity identification and labeling.
2. The system's ability to improve its classification performance through learning.
3. The utility of the system in enhancing operational effectiveness.

Guided by the SCORE framework, component and system level technical performance tests are developed to handle metrics 1 and 2, while system level utility assessments are designed to address the third metric. The quantitative performance tests are accomplished through elemental tests, while the qualitative tests are done through vignette tests.

3.1.1 *Elemental Tests*

This test type was used to measure technical performance at both the component and system levels [4]. Specifically, this test type afforded the designer the ability to place tight controls on the testing environment including modifying specific test variables in order to measure their impact on a technology's performance. The elemental tests that were developed across the ASSIST Phase I and II evaluations include:

- Arabic text translation – This test was designed to evaluate the Arabic text translation ability at both the component and system levels. Component level elemental tests include specific measurements of the technology's ability to 1) Identify Arabic text in an image, 2) Extract Arabic text from an image, and 3) Translate Arabic text to English text. The system level elemental test measured the technology's *start-to-finish* ability from capturing an image of Arabic text and to successfully translating the text into English.
- Face recognition and matching – Likewise, this elemental test evaluated the face recognition technology at the component and system levels. The component level test occurred in the form of an offline evaluation where test images of faces were directly fed into a computer running the matching algorithm and compared against a preloaded watchlist of images. Accuracy measures were calculated based upon the system's output as compared to the ground truth. The system level test evaluated the full hardware/software technology package in a controlled environment by measuring the time and accuracy for the system to capture a person's image and match them against a watchlist.
- Object detection/image classification – This elemental test evaluated these technologies at the system level. The test began with end-users capturing feature/object-laden images of the environment with the evaluation team analyzing their output of the number of objects detected/images classified.
- Shot localization/weapon classification - A system level elemental test was designed to evaluate the accuracy of this technology's ability to detect gunshots, calculate a shot's trajectory, localize a shot's origin, identify the caliber of bullet fired and classify the weapon that fired the shot (see Figure 1 for an example output).
- Soldier state/localization – A system level elemental test was created to assess the ASSIST system's ability to characterize a Soldier's actions within indoor and outdoor environments.
- Sound/speech recognition – A system level elemental test was devised to evaluate the technology's ability to detect specific sounds within the environment.



Idle
Caliber: 5.56 mm
Weapon: M16
Speed: 871 m/s

Figure 1: Shot localization/weapon classification output

Once the technologies and their respective elemental tests were ascertained, the next step was to define the specific metrics and measures. This step also included identifying the influential variables that impact performance, specifically highlighting which variables should be fixed along with those that should be altered during the test(s). More information on this step with respect to the ASSIST evaluations can be found in [2] [4].

It was now time to identify a suitable testing environment for each of these elemental tests. It was determined that the system-level elemental tests would be conducted at a MOUT site at the Aberdeen Proving Grounds in Maryland. The only exception to this was the shot localization/weapon classification test. Since live gunfire was necessary to accurately assess this technology and safety restrictions were in place at the MOUT site, this technology was evaluated at a live fire range adjacent to the MOUT site. Locating a test environment for the component level elemental tests was less taxing since these could be run practically anywhere since they were run on common personal computers (PCs).

Choosing participants was the next step, specifically those that will use the technology (whether it be the members of the end-user population or the technology developers) and those that will indirectly interact with the systems (including those playing roles within the environments). Per DARPA's instructions, Phase I evaluations had the technology developers use/wear their ASSIST systems and shadow the movements of partner Soldiers. This restriction was reduced as both researchers and end-users (Soldiers) used/wore the systems throughout the Phase II evaluations.

Training of these personnel played a critical role in the evaluations. For Phase I that called for the developers to use their own systems, the training consisted of familiarizing these personnel with the scope of the elemental tests, not the technology (since they were the ones that created the systems). However, when the Soldiers stepped in to use the technology during later evaluations, they had to be trained not only on the scope of the elemental tests, but also on how to use the technology. Likewise, the actors in the environment (e.g. the people whose faces were captured to support the face recognition technology, the shooters who fired the weapons to test the shot localization/weapon classification technology, etc.) all had to be trained on their roles.

Additionally, it was necessary to determine how data was to be collected from the ASSIST technologies (and the environment, where necessary). Successfully undertaking this task required that the technology outputs are known (which, according to the SCORE framework, are highlighted as the technologies for evaluation are identified) along with realizing what critical data could be captured from the environment. Data collection can be as simple as measuring the amount of time it takes for the face recognition/matching algorithm to return a match. It can also require more complex actions such as an IET member noting specific actions of the system-wearer into a voice-recorder and then comparing those actions with corresponding times (from their audible notes) to that of a technology system-output log file. For each of the described elemental tests, data collection methods were determined based upon the available output data and the metrics necessary for each evaluation.

Going hand-in-hand with determining the data collection methodology and the required personnel were the scenarios in which the components/systems were tested. These use-case scenarios were developed based upon the expected concept(s) of operation (CONOPS) while keeping in mind the technology's current state of maturity. Specifically, CONOPS is a "formal document that employs users' terminology and a specific, prescribed format to describe the rationale, uses, operating concept, capabilities and benefits of a system" [5]. The challenge in this step is that CONOPS do not often exist for emerging technologies. To surmount this obstacle, the IET developed use-case scenarios with end-user and technology developer input. These test scenarios are presented in great detail in [2] [4].

Going through these SCORE-prescribed steps in order to assess technical performance at the component and system levels produced comprehensive evaluations for the above mentioned ASSIST technologies. SCORE was also applied to develop system level utility assessments in the form of vignette tests.

3.1.2 Vignette Tests

This test type was used to perform System Level Testing – Utility Assessment of the ASSIST technologies [6]. In this case, utility is defined as the value that a technology or piece of equipment provides to an end-user. Utility assessments were uniquely designed given the technology's state of maturity. Typically, a system's utility can still be evaluated prior to its full development where the intent of the assessment is to inform on the system design. Assessments done at the end of a technology's development cycle are intended to validate the value of the system. The former evaluation type is known as formative while the latter is defined as summative.

Since the ASSIST technologies were young in development, these formative vignette tests took the form of several operationally-relevant, mini-mission scenarios where end-users employed the technology in use-case situations to accomplish their mission objectives. Informing the developers about the capabilities of the ASSIST technology became the goal in the design and execution of the SCORE-driven utility evaluations. It should be noted that all of the ASSIST Phase I and II technologies were evaluated under vignette tests with the exception of the shot localization/weapon classification due to safety considerations.

Measures were identified in the form of end-user surveys and semi-structured interviews. The end-users (Soldiers in the case of the ASSIST evaluations) were presented with a suite of survey questions that they answered with respect to their recent experiences with the technology. Furthermore, the Soldiers were interviewed (without the technology developers being present) to gain further insight into what features/capabilities they liked, what they didn't like, and what improvements should be made. The responses were rolled up into technology utility assessments.

The Aberdeen MOUT site presented a small-scale, middle-Eastern-like village where Soldiers frequently train. This test environment provided over a dozen single-story and two-story buildings that challenged the ASSIST technology-laden end-users.

The participants selected to use the technology and to interact with the end-users in the environment were chosen in an identical manner to that of the individuals selected for the elemental tests. Phase I started with the researchers wearing their own technologies and shadowing the Soldiers during the elemental and vignette tests while Phase II put the technology directly on the Soldiers. In both phases, extras/environmental actors were employed to bring about more realism in the vignette test environment. Training for these participants is similar for what was done in support of the elemental tests. When the Soldiers were wearing the technology, they were provided specific training by the research teams so they would be competent in the systems' basic operations.

Some of the data collection methods are already presented in the form of survey instruments and semi-structured interviews. Additionally, several evaluation team members were strategically placed within the environment to observe the Soldiers, the researchers (when they wearing the technology during Phase I), and the extras acting within the environment.

In parallel, the specific vignette mission scenarios were created. After considering the various SCORE-prescribed factors and interviewing subject matter experts, the following mission-scenarios were used throughout the various evaluations in Phase I and II included:

- Presence patrol with deliberate search
- Presence patrol leading to a cordon and search
- Presence patrol and improvised explosive device site reconnaissance
- Assessment of local village with respect to an upcoming election
- Presence patrol leading to checkpoint operations

Prior to the execution of each mission, the Soldiers were briefed on their specific objectives and told to react accordingly to the environment based upon their tactical training. The Soldiers were also reminded of the available ASSIST technologies at their disposal and instructed to use them as they see fit to accomplish their mission objectives.

Using the SCORE framework, elemental and vignette tests were designed and executed to provide the program sponsor with the requested data in addition to informing the researchers on the state of their technologies. The next subsection will show how SCORE has been applied to evaluate another technology.

3.2 TRANSTAC – Phase II

TRANSTAC is another DARPA advanced technology and research program whose goal is to demonstrate capabilities to rapidly develop and field free-form, two-way speech-to-speech translation systems enabling English and foreign language speakers to communicate with one another in real-world tactical situations where an interpreter is unavailable [7]. Several prototype systems have been developed under this program for numerous military applications including force protection and medical screening. The technology has been demonstrated on PDA (personal digital assistant) and laptop platforms. DARPA asked NIST to assess these systems starting in Phase II of this program (another team evaluated these systems during Phase I). Five different research teams presented systems for evaluations during this phase.

The NIST IET applied the SCORE framework to this program because this approach would scale well as the systems continued to mature and DARPA wanted both technical performance and utility assessments of the technology. Specifically, the following test types were conducted during the Phase II evaluations:

1. System usability testing – providing overall scores to the capabilities of the whole system.
2. Software component testing – evaluating components of a system to see how well they perform in isolation.

The IET implemented a two-part test methodology to produce these metrics. Metric 1 was evaluated through the use of structured scenarios within live evaluations, while metric 2 was evaluated through the use of pre-recorded utterances within offline evaluations.

3.2.1 *Offline Evaluations*

The offline evaluations, which represent the *Component Level Testing – Technical Performance* aspect of the TRANSTAC evaluation, were designed to test the TRANSTAC systems with exactly the same set of data so comparison among the systems would truly be “apples-to-apples.” Identical speech utterances, both English and foreign language, were fed into each developer’s system. These utterances were collected from audio recordings from data gathering events. First, an audio file was fed into each system to test the systems’ speech-to-text (S to T) capabilities. Then a text format to test their systems’ text-to-text (T to T) capabilities was fed into each technology. Since the system outputs include translated text and speech, metrics were extracted through comparison of the system outputs to ground truth. A range of metrics including low-level concept transfer and automated metrics were extracted from the offline outputs [8].

Since this evaluation focused on inputting utterances into each development teams’ system, choosing the appropriate test site was trivial. For simplicity, the offline evaluation was conducted at the same site as the live evaluations since there were tighter venue constraints for these tests. Additionally, participant selection and training is very straight forward. An offline evaluation specialist worked with a member of each research team to ensure that each system accepted the offline utterances without incident.

The use-case scenarios under which the utterances (both audible and text speech) were generated stemmed from the supporting data collections (and their respective scenarios) that took place months in advance of the evaluation. This scenario development process began with the IET meeting with the technology’s potential end-users, both English-speaking military personnel and foreign language experts, to determine the representative use-cases in which this type of technology would be most beneficial. Once those situations were established, the IET developed scenarios that were used in the data collections. The data collections brought together English and foreign language speakers to talk/role-play through the data collection scenarios that produced 10 to 20 minute data collection dialogues. Each of the audio dialogues were transcribed and translated. A majority of the data was provided to the developers to train their systems while the remainder was held out by the IET to create the evaluation scenarios. Utterances from the evaluation set were selected to be used in the offline evaluation.

3.2.2 *Live Evaluations*

The live evaluations were performed in two different venues, the lab and the field (both containing facets of *System Level Testing – Technical Performance* and *System Level Testing – Utility Assessment*) and were conducted with structured scenarios. This scenario type provided a set of questions to the English speaker that they needed to find answers, while the foreign language speaker was given the answers to those questions in paragraph format. A dialogue occurred between the two speakers and the number of questions that the English speaker was able to get answered were noted. In addition, surveys were provided to the English and foreign language speakers to gauge their perception of the TRANSTAC systems.

Lab evaluations were designed to test the TRANSTAC systems in an idealistic environment, with no background noise and the participants being stationary. The TRANSTAC systems were placed on a table as opposed to being worn by the speakers. This idealistic environment gave the IET and the developers an idea of the best that the systems can do at this stage in their development.

The purpose of the field evaluations was to test the TRANSTAC systems in a more realistic environment. This included well-controlled background noise, the English-speakers carrying the TRANSTAC systems, and both the English and foreign language speakers being mobile during the evaluation.

Twenty structured scenarios (ten in the lab and ten in the field) were designed to foster the evaluation dialogues. These scenarios were derived from the same held back scenarios that the offline scenarios originate.

The system users for these evaluations were chosen to be potential TRANSTAC system end-users including both English-speaking military personnel and representative foreign language speakers. Training and preparation of these individuals was critical. These individuals had to be both trained on the proper usage of the TRANSTAC systems, but also had to be educated on the procedures and flow of the structured scenarios. This training was done sequentially to enable the IET the ability to isolate and address any areas of concern.

Selecting a site for these evaluations required the consideration of numerous factors including a location that could support the offline, lab, and field evaluations, a spot that could accommodate 50+ personnel, a site that was available for six consecutive days, etc. Ultimately, the NIST campus was selected after extensive exploration.

3.3 Initial Findings using SCORE

The ASSIST Phase I and II and TRANSTAC Phase II evaluations were successful events that provided DARPA with the necessary and detailed results desired by the programs. SCORE is viewed as a contributor to this success due to the extensive nature in which it laid out these evaluations. Following its prescribed steps and addressing each evaluation component ensured that comprehensive and relevant evaluations were generated. The following section will show how SCORE has evolved to produce innovative evaluations as the ASSIST and TRANSTAC programs further advance.

4. EVOLUTION of the FRAMEWORK

Both the ASSIST and TRANSTAC programs have since moved into Phase III. To date, two ASSIST Phase III evaluations have been performed while a single TRANSTAC Phase III evaluation has already occurred with each program having one more Phase III evaluation to go. Since these programmatic goals have changed from the previous phases, the SCORE framework has evolved to produce the desired metrics. One major innovation is the addition of a fourth evaluation goal type, described below:

- *Capability Level Testing – Utility Assessments* – This evaluation group is proposed to assess the utility of an individual capability (where the complete system is made up of multiple capabilities), where utility is defined as the value the application provides to the system end-user (just as it is *System Level Testing – Utility Assessments*). The benefit of this evaluation type is that specific capability utility and usability to the end-user can still be addressed even when the system and user-interface are still under development.

This goal type can be inserted into the tiered approach either after the *Component Level Testing – Technical Performance* or the *System Level Testing – Technical Performance* goal types.

Each of the evaluation elements described in section 2.1 are applied to this new goal type. This new SCORE addition will be presented within the following discussion of the ASSIST Phase III evaluation design whereas further applications of SCORE will be discussed in TRANSTAC Phase III evaluation plan.

4.1 ASSIST – Phase III

As the ASSIST program moved into Phase III, the program evaluation focus was altered to place more emphasis on end-user utility assessments as opposed to technical performance. With the technologies further along in their development cycles (as compared to their status in earlier phases), it was becoming more important to gain insight into the end-users' value of specific capabilities. In addition to emphasizing utility assessments, the program is now more focused on real-time capabilities as opposed to those that support after-mission reporting. Three separate research teams produced Phase III evaluation technologies that included the following capabilities:

- Face recognition/matching
    - o Face image collection using commercial off-the-shelf (COTS) hardware
    - o Face image matching displayed on COTS wearable interface
- Real-time information collection and sharing
    - o Automatic capture of image, audio, and GPS data
    - o End-user viewing own-captured data on wearable, COTS visual display (see Figure 2)
    - o Transmit/receive image and GPS data to/from other ASSIST units
- After-mission reporting
    - o Retrieving mission data at specific times and locations
    - o Locating field-marked significant actions
    - o Observing soldier state analytics

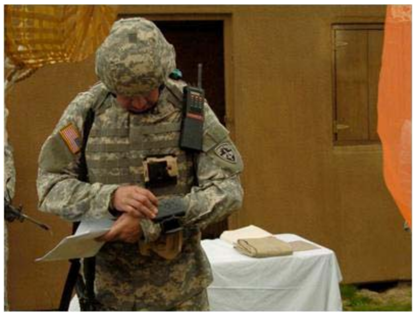


Figure 2: Soldier using real-time data collection ASSIST system

To satisfy the program goals, only technical performance evaluations were designed for the face recognition/matching technology at both the component and system levels. This took the form of elemental tests, similar to those outlined in section 3.1.1. Additionally, system level utility assessments were collected for the real-time information sharing and after-mission reporting technologies through additional vignettes (comparable to those presented in section 3.1.2.).

However, the need to gather further utility assessments, especially of the face recognition/matching technology which was not evaluated in the vignette tests during this phase, spawned the SCORE evaluation goal type of *Capability Level Testing – Utility Assessment.* This inspired the development of task tests whose intent was to assess end-user utility of specific capabilities within the various ASSIST technologies.

After determining the objective of the task evaluations, the IET continued down the path of identifying the remainder of the SCORE evaluation elements by identifying the necessary measures and metrics. The measures extracted from this test include IET observer notes (made while following the end-users with the technology during the tasks) along with surveys presented to the end-users at the conclusion of each task (similar to those given at the end of the vignette tests). The data collection methods used to gather the observer notes include the use of hand-held PDA note-taking devices while the surveys were administered via PC. The survey results and observer notes were combined to produce the necessary metrics (similar to what was done to produce the metrics from the vignette tests).

These task tests, in addition to the elemental and vignette tests, took place at the same Aberdeen MOUT site that supported the Phase I and II evaluations. Multiple participants were required for the task tests. Soldiers, the ultimate end-users, were selected to use/wear the systems throughout the task tests. For the task tests (like the other test types), training was a critical component. Specific time was set up for the research teams to brief the Soldiers on their technology along with allowing them an opportunity to have hands-on practice with the various systems. Additionally, training time was also allocated for the Soldiers to become competent with the specific task tests (both, the test objectives and flow). Prior to the Soldiers running these tasks, they were briefed on the specific task objectives (both with respect to using the technology and the tactical goals). Following the briefing, the end-users and the IET practiced each of the task runs (without the technology) to ensure everyone was competent with the tests when it became time to run them with the technology.

The task test scenarios were created in parallel to addressing the numerous steps presented above. Tasks were developed to specifically address all of the Phase III technology capabilities including the following:

- Street observation and interaction – This task was developed to specifically test real-time image sharing across multiple ASSIST systems.
- Presence Patrol – This task was designed to evaluate personnel tracking, GPS positioning and map annotation capabilities.
- Insurgent Surveillance – This test was created to assess the capability of image and map transfer between the laptop-based systems and ground-based wearable ASSIST technologies.
- Insurgent Surveillance and Ambush – This task was created to test the ASSIST technology's ability to calculate soldier state analytics.
- Base/entry checkpoint – This task was developed to test the face recognition/matching system's ability to capture images in the field and present matches in real-time on the system-wearer's personnel interface.

These tasks were designed to be between 10 to 15 minutes in length where each was run twice. The runs were also set up to have three end-users use the relevant ASSIST technology with an emphasis on the specified capabilities. Two Soldiers used the portable, wearable technology while the other user interacted with the laptop-based system.

Addressing each one of the SCORE framework elements with respect to the task tests enhanced the effectiveness of this series of evaluations at the most recent ASSIST events. Comprehensive utility assessments were collected from the task tests which enabled the IET to produce an extensive picture of the current state of the ASSIST technologies when combined with the elemental and vignette test data.

This additional *Capability Level Testing – Utility Assessment* is an advancement in the SCORE framework. Additional improvements will be shown in the following section discussing the latest phase of the DARPA TRANSTAC program.

4.2 TRANSTAC – Phase III

Phase III of the DARPA TRANSTAC program continues to present the same overall evaluation objectives as presented in Phase II. Additionally, this phase brought about additional technical performance and utility assessments of several specific TRANSTAC technologies including a ruggedized, portable hardware platform (known as the Lynx system) and the systems' ability to handle the translation of names, streets, and places (simply stated as "names" throughout the rest of this paper) from a specific foreign language to English.

Keeping these goals in mind, offline and live formats (lab venue, only) were conducted similar to those run in Phase II to accomplish the primary evaluation goals. The SCORE framework played a critical role in defining the testing scopes for evaluating the Lynx system and the systems' capability to address names.

The Lynx evaluation was designed under the *System Level Testing – Technical Performance* and *System Level Testing – Utility Assessment* evaluation goal types. The design of this evaluation closely mirrored that of the live lab evaluations. Recall that the main TRANSTAC systems were evaluated with both speakers sitting at a table interacting with the laptop-based system which was placed on the table (as opposed to being worn). To that end, the Lynx systems were evaluated in a similar manner where they were placed on a table where the English speaker sat on one side of the table and the foreign language speaker on the opposite. The Lynx test tasked the speakers with transferring as many concepts as possible within a ten minute timeframe while adhering to the structured scenario format. For the sake of comparison, the same structured scenarios that were used in the main evaluation were selected for the Lynx evaluation. As in the main test, the evaluation team was able to extract technical performance metrics through the number of concepts transfer. Additionally, the end-users were administered specific surveys to assess their utility of the Lynx technology.

Because the Lynx platform was different from that of the laptop-based systems, additional training was provided to the end-users before this evaluation. This was particularly important so that the end-users did not confuse this system's operation with that of the technology they had used earlier (the main laptop system evaluations were conducted immediately prior to the Lynx system testing).

The names capability was evaluated under the *Component Level Testing – Technical Performance* and *Capability Level Testing – Utility Assessment* evaluation goal types. This test was conducted in both the live lab and offline settings and used the main evaluation laptop-based platforms. The only other similarity to the main live lab evaluations include the fact that the speakers were sitting across from one another at a table and did not have to wear the system.

Since the goal of this evaluation was to isolate the systems' ability to translate names, the SCORE elements directed the IET to design unique scenarios to support both the offline and live lab venues. This specialized scenario design stemmed back to the data collection scenarios. Three unique, names-laden scenarios were created as scripted dialogues and recorded by unique speaker-pairs. These dialogues were crafted such that there was at least one name in each foreign language utterance where it was noted whether this name appeared in the names lexicon (a list of names that the research teams have access) or if it did not along with whether each name was unique (the name can only mean a name) or whether it was a "double" (the name can also mean an object, etc). This recorded data was used to create the offline names evaluation set where all of the recordings were kept by the IET (no names data was released to the developers since the intent was to prevent the out of lexicon names from being known by the researchers ahead of time). The scenarios used in the live names evaluations were identical to those scripted ones used in the names data collections.

The offline names evaluation ran similarly to that of the main offline evaluation. Specific utterances are selected and fed directly into the TRANSTAC systems. However, the measures and metrics from this test focused on how the systems specifically handled the translations of the names.

The live names evaluation ran in a different manner than that of the live main (lab) evaluation. The speakers were provided with the scripted names scenarios (as opposed to the standard structured scenarios) and instructed to read them verbatim. The English speaker began each utterance by stating the number of the utterance they were about to read (which alerts the foreign language speaker to the current utterance) and then spoke the utterance. Since the focus of this evaluation was on the translation of names from the foreign language into English, the English speaker did not speak into the TRANSTAC system. After hearing the English utterance, the foreign language speaker responded with their scripted utterance which they spoke into the TRANSTAC system. If the English speaker was able to understand the name that was communicated, they noted that and moved on to the next utterance. If the English speaker was unable to ascertain a name from the TRANSTAC output, then they were able to rephrase their utterance in any manner they saw fit. Likewise, the foreign language speaker, upon hearing the English speaker rephrase their utterance, rephrased theirs accordingly to convey the desired name. The output of this evaluation produced both technical performance and utility assessment data. This took the form of measuring the number of names successfully transferred and collecting survey responses from the end-users regarding their specific names interactions.

The SCORE framework was successfully employed to address additional evaluation goals including the Lynx system and names translation capabilities. Likewise, the framework further evolved to address progressing needs in the ASSIST program.

5. FUTURE EFFORTS

SCORE is still being used to design the remaining ASSIST and TRANSTAC Phase III evaluations which will both take place before the end of the calendar year. If these programs continue, it is envisioned that SCORE will be used to design their successive evaluations.

The SCORE framework is applicable to domains beyond emerging military technologies and those solely dealing with intelligent systems. Personnel at NIST are applying the SCORE framework to the virtual manufacturing automation competition (VMAC) and the virtual RoboRescue competition (within the domain of urban search and rescue). Their intent is to develop elemental tests and vignette scenarios to test complex system capabilities and their component functions. Likewise, personnel in NIST's construction metrology group have expressed interest in the SCORE framework with respect to designing evaluations within the automated construction domain.

It is envisioned that SCORE will be applied to a broad range of technologies, both to design evaluations of emerging components and systems along with enhancing evaluation procedures of pre-existing technologies. This framework is highly adaptable and capable of meeting most any evaluation requirement.

6. CONCLUSION

SCORE has proven to be an invaluable evaluation design tool of the NIST IET and was the backbone of eight (five for ASSIST and three for TRANSTAC) evaluations. Further, it is expected to play a critical role in the remaining Phase III ASSIST and TRANSTAC evaluations. The NIST IET will continue to apply the SCORE framework in future evaluations (including those outside of the military community) and will support other members in the technology evaluation community who wish to leverage it.

7. DISCLAIMER

Certain commercial products and software are identified in this paper in order to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the products and software identified are necessarily the best available for the purpose.

8. ACKNOWLEDGMENTS

The authors would like to acknowledge the DARPA ASSIST/ TRANSTAC program manager, Dr. Mari Maeda, and the members of the NIST IET for their continued support.

9. REFERENCES

[1] Steves, M. and Scholtz, J. 2005. A Framework for Evaluating Collaborative Systems in the Real World. In *Proceedings of the 48th Hawaii International Conference of System Sciences* (HICSS-38), Hawaii.

[2] Schlenoff, C., Steves, M., Weiss, B.A., Shneier, M., and Virts, A. 2006. Applying SCORE to Field-based Performance Evaluations of Soldier Worn Sensor Technologies. *Journal of Field Robotics*, Volume 24, (8-9), pp. 671-698.

[3] Schlenoff, C., Weiss, B.A., Steves, M., Virts, A., Shneier, M. and Linegang, M. 2006. Overview of the First Advanced Technology Evaluations for ASSIST, In *Proceedings of the 2006 Performance Metrics for Intelligent Systems (PerMIS) Conference,* Gaithersburg, Maryland.

[4] Weiss, B.A., Schlenoff, C., Shneier, M., and Virts, A. 2006. ASSIST: Technology Evaluations and Performance Metrics for Soldier-Worn Sensor Systems. In *Proceedings of the 2006 Performance Metrics for Intelligent Systems (PerMIS) Conference*, Gaithersburg, Maryland.

[5] CONOPS, http://en.wiktionary.org/wiki/CONOPS, (Definition retrieved on 07/10/08)

[6] Steves, M. P. 2006. Utility Assessments for ASSIST Systems. In *Proceedings of the 2006 Performance Metrics for Intelligent Systems (PerMIS) Conference*, Gaithersburg, Maryland.

[7] Weiss, B.A., Schlenoff, C., Sanders, G., Steves, M., Condon, S., Phillips, J., and Parvaz, D. 2008. Performance Evaluation of Speech Translation Systems, In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference*. Marrakech, Morocco.

[8] Sanders, G., Bronsart, S., Condon, S., and C. Schlenoff. 2008. Odds of successful transfer of low-level concepts: A key metric for bidirectional speech-to-speech machine translation in DARPA's TRANSTAC program. In *Proceedings of LREC 2008.*