Proceedings of the ASME International Design Engineering Technical Conferences &
Computers and Information in Engineering Conference
IDETC/CIE 2010
August 15 – 18, 2010, Montreal, Quebec, Canada

DETC2010-28928

THE MULTI-RELATIONSHIP EVALUATION DESIGN FRAMEWORK: DESIGNING TESTING PLANS TO COMPREHENSIVELY ASSESS ADVANCED AND INTELLIGENT TECHNOLOGIES

Brian A. Weiss
National Institute of Standards and Technology (NIST)
Gaithersburg, Maryland, USA

Linda C. Schmidt
University of Maryland
College Park, Maryland, USA

Harry A. Scott
NIST
Gaithersburg, Maryland, USA

Craig I. Schlenoff
NIST
Gaithersburg, Maryland, USA

ABSTRACT

As new technologies develop and mature, it becomes critical to provide both formative and summative assessments on their performance. Performance assessment events range in form from a few simple tests of key elements of the technology to highly complex and extensive evaluation exercises targeting specific levels and capabilities of the system under scrutiny. Typically the more advanced the system, the more often performance evaluations are warranted, and the more complex the evaluation planning becomes. Numerous evaluation frameworks have been developed to generate evaluation designs intent on characterizing the performance of intelligent systems. Many of these frameworks enable the design of extensive evaluations, but each has its own focused objectives within an inherent set of known boundaries.

This paper introduces the Multi-Relationship Evaluation Design (MRED) framework whose ultimate goal is to automatically generate an evaluation design based upon multiple inputs. The MRED framework takes input goal data and outputs an evaluation blueprint complete with specific evaluation elements including level of technology to be tested, metric type, user type, and, evaluation environment. Some of MRED's unique features are that it characterizes these relationships and manages their uncertainties along with those associated with evaluation input. The authors will introduce MRED by first presenting relationships between four main evaluation design elements. These evaluation elements are defined and the relationships between them are established including the connections between evaluation personnel (not just the users), their level of knowledge, and decision-making authority. This will be further supported through the definition of key terms. An example will be presented in which these terms and relationships are applied to the evaluation design of an automobile technology. An initial validation step follows where MRED is applied to the speech translation technology whose evaluation design was inspired by the successful use of a pre-existing evaluation framework. It is important to note that MRED is still in its early stages of development where this paper presents numerous MRED outputs. Future publications will present the remaining outputs, the uncertain inputs, and MRED's implementation steps that produce the detailed evaluation blueprints.

INTRODUCTION

Innovative technologies, including those designated as intelligent systems or deemed to have robotic elements, are regularly developed and for use across a wide range of areas such as manufacturing, law enforcement, military, urban search and rescue, and autonomous vehicles. Evaluating these technologies is critical to inform designers during the design process and validate performance of final systems. The human-robot interface (HRI) or human computer interaction (HCI) is the fundamental commonality of all [14] [15] [16]. No matter the level of intelligence of a system, there is always a human-in-the-loop. The human operator may be controlling all system functions, observing the robot's behavior, or exerting varying levels of control in between these two extremes.

Autonomous ground vehicle technologies, including their intelligent control architectures, automated positioning and mapping systems, are advanced technologies that have evolved over the past decades. These technologies have spurred the development and implementation of highly complex evaluations, a majority of them before the final systems are deployed [1] [4] [17].

Another area of advanced technology is Urban Search and Rescue (US&R) and bomb disposal robotic systems. To date, an extensive array of tests have been produced, conducted and refined to evaluate US&R and bomb disposal robots across a range of operational scenarios [8] [9] [10]. This array includes test groupings designed to evaluate different system capabilities such as mobility, directed perception, grasping dexterity, visual acuity, etc. One such collection of tests used to evaluate US&R systems includes random stepfield pallets to challenge the mobility across varying terrains [7]. Stepfield pallets were developed to represent complex terrain or rubble that is describable, reproducible, and repeatable for testing robotic systems. The robots are both tested against the stepfield pallets directly to evaluate their mobility or used as a secondary test to see their impact on system performance while a robot is trying to accomplish another task (such as manipulate an object). These tests are governed by specified variables which dictate the test conditions. In this case, test variables include human operators and stepfield arrangements. These conditions will change as the evaluation goals update.

Many government and private institutions have invested in research and development of frameworks to effectively and comprehensively assess the performance of intelligent systems. Nearly all of the frameworks have been suitable to evaluate their given technologies and achieve program-specific goals, but no single framework has been identified as being suitable to assess both quantitative and qualitative performance across a wide-range of virtual and physical systems that include both human-controlled and autonomous functions.

The National Institute of Standards and Technology (NIST[2]) developed the System, Component, and Operationally-Relevant Evaluation (SCORE) framework to effectively evaluate numerous advanced and intelligent technologies at multiple levels [11]. The SCORE framework has been successfully applied over a dozen evaluations. Its success has generated extensive quantitative and qualitative data that have proven beneficial to the system developers, evaluation designers, potential end-users, and funding sponsors.

The primary author, also a co-creator of SCORE, will draw from the success of SCORE to develop an evaluation framework that will automatically produce evaluation blueprints (test plans). The ultimate goal of the MRED framework is to produce an evaluation design framework that will take inputs from three specific groups, each with their own levels of uncertainty, and output an evaluation blueprint that clearly specifies all aspects of the test event(s). For this work, the authors define an evaluation blueprint as a detailed test plan that states the levels and values of the test variables and how they will be combined to set up and implement the test. The blueprint also specifies the class(es) of metrics to be collected which would either include quantitative and/or qualitative data.

This paper will discuss the following: the SCORE framework will be presented including its background and specific evaluations it has supported; the authors' proposed Multi-Relationship Evaluation Design (MRED) framework will be discussed; an example of the design of a vehicle evaluation will be used to show how these terms and relationships are implemented; MRED will be applied to an evaluation design previously inspired by SCORE to validate MRED's applicability; and plans to further develop and enhance MRED will be discussed.

It is important to note that this paper will only extensively cover specific evaluation frameworks' output elements. The remaining blueprint elements along with the framework's inputs will be covered in future work.

OTHER FRAMEWORKS

Many test planning systems have been devised to evaluate complex emerging and intelligent systems. An evaluation framework was developed to assess mobile robots for planetary exploration across relevant terrains, but this did not consider the element of human interaction and was designed with a mission-specific emphasis [19]. Another evaluation framework has been designed specifically to assess intelligent algorithms where the test output yields extensive and informative quantitative data [5]. This framework has been very successful in capturing technical performance in the virtual world, but hasn't been applied to capture qualitative feedback from human users nor assessing physical implementations. The United States Army has evaluated network-enabled systems, although this has required the usage of multiple test and evaluation techniques as opposed to implementing a single unified framework [6]. Specifically, four strategies have been employed where each is capable of designating evaluations at specific technology levels which is both an advantage and disadvantage; each strategy excels at designing an evaluation at its specific technology level, but all four must be applied in order to produce comprehensive assessments. Additional test methodologies have been developed for use by the Army including the Unmanned Autonomous System Testing (UAST) intended to measure the intelligence of unmanned autonomous systems [20]. The UAST framework is capable of evaluating both virtual and physical systems at both system and sub-system levels, but its current work has yet to focus on producing qualitative measures specified by the users and only specifies pass/fail measures based upon mission tasks.

The SCORE framework was conceived "around the premise that intelligent systems must be evaluated at the component level, the system level, and in operationally-relevant environments" [11]. The SCORE framework must be comprehensive and adaptable to apply to technologies at many different points during their development cycle [18]. SCORE has been successful in enabling evaluation designers to identify

[2] 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.

the most useful blueprints for evaluating various intelligent systems. MRED will draw from the success of SCORE to produce a framework that will automatically output evaluation blueprints given goal inputs. MRED will also recognize the relationships and interdependencies among evaluation elements and address the uncertainty from the evaluation input and how it impacts the blueprint. Since MRED leverages elements from the SCORE framework, SCORE will be discussed in further detail.

LEVERAGING THE SCORE FRAMEWORK

The SCORE framework specifies a set of elements that define performance evaluations of intelligent systems. SCORE produces an evaluation strategy that is able to generate both technical performance and end-user utility assessments within a host of test venues [18]. Technical performance focuses on quantitative measurements, while utility assessments include qualitative judgments of the technology during the test event. SCORE's utility assessment evaluation type measures usability, effectiveness, and user attitude towards the technology.

To use SCORE, designers determine their goal(s) with respect to the system, components, and/or specific capabilities to be addressed. Evaluation goal types is a specific piece of SCORE that MRED will leverage in its framework.

Evaluation Goal Types

Five evaluation goal types are specified by the SCORE framework. Goal types are combinations of the technology level and desired metrics [12] [13] [18]. Note that a capability is a behavior that is produced either from a single component or multiple components working together. Depending upon their relationships, capabilities and components may be separated for evaluation.

- *Component Level Testing – Technical Performance* – Evaluation type breaks down a system into components in order to separate the subsystems that are essential for system functionality and can be designed or altered independently of other components.
- *Capability Level Testing – Technical Performance* – Evaluation type requires the identification and isolation of specific capabilities from overall system behavior to the measure the individual capabilities' contribution to technical performance.
- *System Level Testing – Technical Performance* – Evaluation type targets a full system assessment where environmental variables can be isolated and manipulated to capture their impact on system performance.
- *Capability Level Testing – Utility Assessments* – Evaluation type assesses the end-users' utility of a specific capability where the complete system's behavior is composed of multiple capabilities. In this instance, the SCORE framework defines utility as the value the application provides to the end-user.
- *System Level Testing – Utility Assessments* – Evaluation type focuses on the end-users' utility of the entire system.

For each of the evaluation goal types, SCORE specifies numerous elements that must be defined in the blueprint [11] [18] [23]:

- *Identification of the system, component, or capability to be assessed*
- *Definition of the goal, objective(s), metrics, and measures*
- *Specification of the testing environment*
- *Identification of the personnel*
- *Specification of the personnel training*
- *Specification of the data collection methods*
- *Specification of the use-case scenarios*

It is critical to identify these evaluation elements in order to produce fair and meaningful results. The existence of a number of evaluation goal types and elements leads to a vast number of potential testing blueprints. In the absence of an assessment planning tool, evaluation designers have to determine the most appropriate of these options based on experience alone.

SCORE Successful Applications

Since 2005, the SCORE framework has successfully guided 14 evaluations or competitions by enabling evaluation designers to prescribe specific evaluation goal types across a range of evaluation elements. SCORE has been used as the backbone of six tests (each test event spanning multiple weeks) focused on the quantitative and qualitative evaluation of soldier-worn sensor systems. The goal of this project is to provide warfighters with real-time data collection and information sharing technology within the battle-space along with enhanced after-mission data-mining and display capabilities [11] [13] [18] [22] [25]. Based upon the evaluation goals, SCORE prescribed a range of tests that included all five of the evaluation goal types.

Additionally, SCORE served as the framework to generate evaluation blueprints for five week-long tests for military-supported developmental speech-to-speech translation systems. These technologies are intended to provide military personnel with two-way, free-form, spoken language translation devices for use in various tactical situations [12] [19] [22] [23] [24]. SCORE successfully dictated the design and implementation of over six unique evaluations across all five evaluation goal types to effectively yield the metrics required by the program.

The SCORE framework has also supported the design of the Virtual Manufacturing Automation Competitions (VMAC) whose intent is to test a range of algorithms within a simulated manufacturing world [2]. To date, SCORE has contributed to three VMACs including competitions on both the national and international stages.

The authors seek to develop the MRED framework so that it will be applicable across a wider range of technologies by exploiting relationships among evaluation elements and addressing uncertainties.

MRED EVALUATION DESIGN FRAMEWORK

The Multi-Relationship Evaluation Design (MRED) framework is presented by first highlighting the critical inputs into the planner and the nature of the output "evaluation blueprint". Three distinct input groups are denoted as essential information for MRED as shown in Figure 1. Each group is described separately.

Input Group 1 – Stakeholders

There are six personnel groups that all have a vested interest in a technology's evaluation and that could have influence over the design of a technology evaluation. Each of these groups has their own motivations regarding the technology evaluation along with certain levels of uncertainty in their preferences. The six personnel groups are:

- *Buyers* – The personnel group who are willing to purchase the technology.
- *Users, Potential Users* – The personnel group that are already using the technology or who are the target user group. Individuals from this group may or may not be *Buyers*.
- *Evaluation Designers* – The personnel group who design the technology evaluation by determining the appropriate MRED inputs.
- *Evaluators* – The personnel group that implements the evaluation design. The *Evaluators* may also be the *Evaluation Designers*.
- *Sponsors/Funding Sources* – The personnel group who are funding the technology development and/or evaluation.
- *Technology Developers* – The personnel group that is responsible for designing and building the technology.

Input Group 2 – Technology State

This represents the anticipated maturity and functionality of the technology when it is to be tested. Has the technology gone through previous testing where the output test data has been used to iterate upon the design? Does the technology currently have all of its intended functionality or is test feedback required before the system can be finalized. These questions provide some insight as to whether or not the evaluation will be *formative* (intended to inform on a technology's design while it's still in development and not fully mature) or *summative* (intended to validate the final design of a technology) [18].

Input Group 3 – Available Resources for Testing and Analysis

The final input group is comprised of specific types of material, manpower and technology to be included in the testing exercise. Resource availability (or lack thereof) and their limitations can have a significant impact on the final evaluation design:

- Personnel – This includes those individuals that will use the technology during the test(s), those that will indirectly interact with the technology during the test, those that will collect data during the test, and those that will analyze the data following the test(s).
- Environment – This includes the physical test venue, supporting infrastructure, artifacts and props that will support the test.
- Data Collection Tools – This includes the tools, equipment, and technology that will collect quantitative and/or qualitative data during the test(s).
- Data Analysis Tools – This includes the tools, equipment and technology capable of producing the necessary metrics from the collected evaluation data.

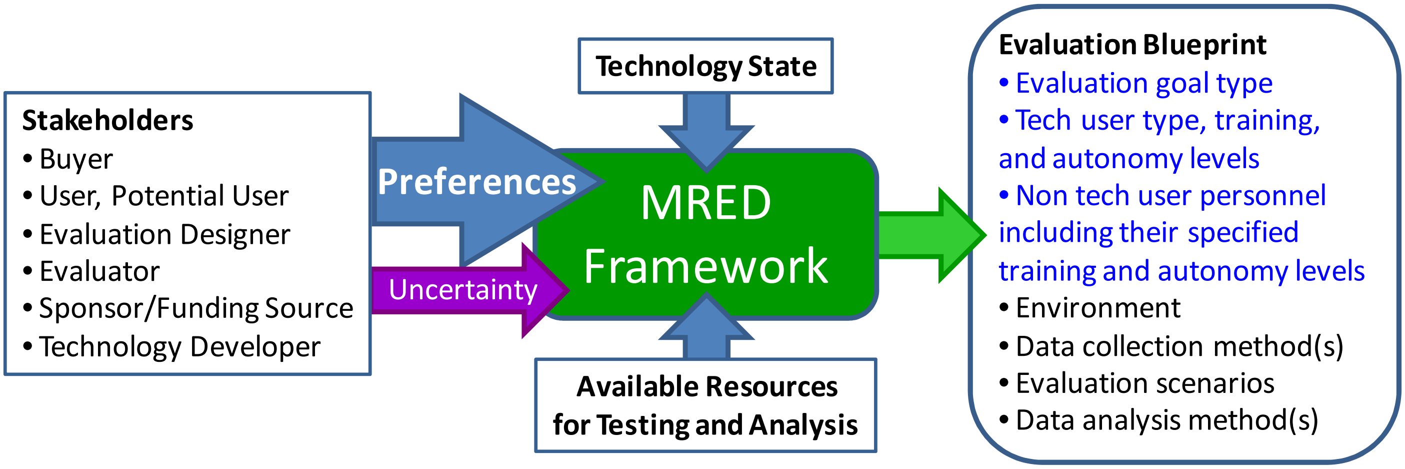


Figure 1 – Input (Groups 1 to 3) and Output (Evaluation Elements) into the MRED Framework

Specific components of Figure 1 will be covered in the remainder of this section including evaluation goal types (through pairings between technology levels and metrics) and personnel including their specific levels of training and decision making authority. This discussion begins by first proposing several relationships between critical evaluation elements. The correlation between technology levels, metric types, technology users and the evaluation environments will be presented, followed by the connection between evaluation personnel (not just technology users), their knowledge levels and their decision-making autonomy.

Before presenting these relationships, it is important to define some of the key terms relating to the evaluation blueprint elements. The authors define these terms with respect to this specific work so as not to be confused with their usage in other areas.

Definition of Key Terms in Blueprint

Key terms in the evaluation blueprint are presented in the following subsections that contribute to the evaluation blueprint. Some of the key terms are explicit in the blueprint, such as personnel and environment. The remaining terms are implicitly present, such as evaluation goal types being comprised of technology levels and metric types.

Technology Levels A technology or system is made up of constituent components and therefore can be evaluated at these multiple levels. The terms relating to technology levels are defined as follows:

- *System* – Group of cooperative or interdependent components forming an integrated whole intended to accomplish a specific goal.
- *Component* – Essential part or feature of a system that contributes to the system's ability to accomplish a goal(s).
- *Sub-Component* – Element, part or feature of a *Component*.
- *Capability* – A specific ability of a technology where a *System* is made up of one or more *Capabilities*. A *Capability* is provided by either a single *Component* or multiple *Components* working together.

Metric Types Evaluations are capable of collecting two unique types of metrics. Before defining the two metric types, it is important to define metrics and measures in the context of this work [11] [23].

- *Measures* – A performance indicator that can be observed, examined, detected and/or perceived either manually or automatically.
- *Metrics* – The interpretation of one or more contributing elements, e.g. measures that correspond to the degree to which a set of attribute elements affects its quality.

The two metric types are:

- *Technical Performance* – Metrics related to quantitative factors (such as accuracy, precision, time, distance, etc). These metrics may be required by the program sponsor, to meet *User* expectations, inform the *Technology Developers* on their design, etc.
- *Utility Assessments* – Metrics related to qualitative factors that gauge the condition or status of being useful and usable to the target end user community.

Personnel Numerous individuals and groups are necessary in order to execute an effective evaluation. They can be classified into two categories: primary (direct interaction) technology users and secondary (indirect interaction or evaluation support). These primary technology users are defined as:

- *Tech User* – Individuals that directly interact with the technology during the evaluation. These individuals receive any training necessary to use the technology and are responsible for engaging/disengaging the technology's usage during the test event. There are multiple classes of *Tech Users* defined below. *Tech Users* are usually the predominant source of qualitative data when the evaluation goal(s) include capture of utility assessments.
    - *Tech User: End-User* – Individuals that are the intended users for the technology. Depending upon the level and extent of the evaluation, all, some, or none of the *Tech Users* will be from the *End-User* class.
    - *Tech User*: *Trained User* – Individuals selected to be *Tech Users*, but who are not *End-Users*. They receive all of the necessary training that *End Users* would receive, but they do not have the operational background or experiences of the *End Users* within the technology's targeted use case environment.
    - *Tech User: Tech Developer* – Members of the research and development organization that developed the technology under evaluation. This category of personnel does not have the operational background or experiences of an *End-User*, but they are intimately familiar with the technology's operations. *Tech Developers* may be *Tech Users* depending upon the level and extent of the evaluation. If so, then they may not require the full training complement.

The relationships among these primary technology user groups along with a secondary technology user (*Team Member)* are highlighted below in Figure 2.

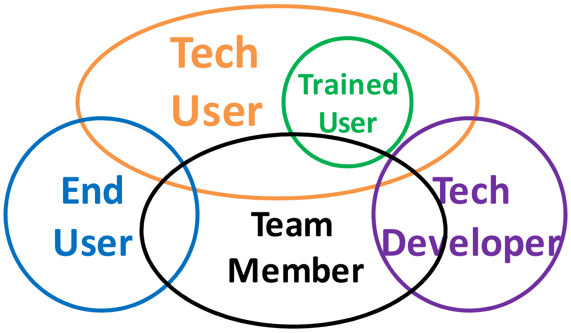


Figure 2 - Relationship among the potential technology users

The secondary personnel, those that indirectly interact with the technology and/or support the evaluation, fall into the following three categories:

- *Team Member* – Individuals that work with *Tech Users* during the evaluation as they would to realistically support the use-case scenario that the technology is immersed. *Team Members* may or may not be in a position to indirectly interact with the technology during the evaluation, but they are often in a position to observe a *Tech User's* interactions with the system. Depending upon the evaluation, *Team Members* may be requested to provide their perceptions of a *Tech User's* utility of the technology along with their level of situational awareness while using the technology, etc. *Team Members* may also be designated as secondary users in real situations meaning they would have some technology training in these cases.
- *Participant* – Individuals that indirectly interact with the technology during an evaluation. Typically, *Participants* are given specific tasks to either interact with the *Tech Users* and/or with the environment, but not with the technology (unless directed to do so by a *Tech User*).
- *Evaluator* – Members of the evaluation team present within the test environment that tasks the *Participants* and/or captures data, but does not interact with the technology. Depending upon the test, the *Evaluator* may interact with the *User* to capture data.

Environment The venue in which the evaluation occurs can have an impact on the data since the environment can influence the behavior of the personnel as well as restrict which levels of a technology can be evaluated. Three distinct environments are identified below:

- *Lab* – Controlled environment where test variables and parameters can be isolated and manipulated to determine how they impact system performance and/or the *Tech Users'* perception of the technology's utility.
- *Simulated* – Environment outside of the *Lab* that is less controlled and limits the evaluation team's ability to control influencing variables and parameters since it tests the technology in a more realistic venue.
- *Actual* – Domain of operations that the system is designed to be used. The evaluation team is limited in the data they can collect since they cannot control environmental variables (doing so would make this a *Simulated* environment).

Knowledge Levels The *Tech Users* and *Participants* involved in the evaluation have various levels of knowledge about aspects of the system and testing conditions within two specific areas. The levels are defined as:

- *Operational Knowledge* – The level of practical information and experience an individual has about the *Actual* environment, the intended use-case situations for the technology and other pre-existing technologies that the technology under test leverages and/or supports. Varying levels of *Operational Knowledge* can be attained through real-world experience, repetitive training, trial and error exercises, etc.
- *Technical Knowledge* – The level of information and experience an individual has about the technology itself and how it should be employed to maximize success.

Autonomy Levels Additionally, the *Tech Users* and *Participants* within the evaluation have a range of Decision-Making (DM) autonomy. Autonomy scope and their levels are set by MRED for each evaluation. Personnel could be fully restricted in their decision-making (i.e., no DM Autonomy), which leads to scripted actions. Alternatively, personnel may have unbounded decision-making authority where each participant is free to exercise their judgment given their various knowledge levels. Specifically, there are two types of DM Autonomy which are defined below:

- *DM Autonomy – Technical* – This refers to the level of authority that the *Tech Users* have in operating the technology. Depending upon the specific evaluation, *Tech Users* could be instructed to only use certain features of a technology to being told that they may use any or all of its features as they see fit.
- *DM Autonomy – Environmental* – This refers to the level of authority that the *Tech Users* and *Participants* have in interacting with each other and the environment.

Now that all of the critical terms are defined, relationships between the variable groups can be explained.

Relationships Between Levels, Metrics, Users, And, Environments

The first relationship to be presented is that between the technology levels, metric types, tech users, and evaluation environments. Individually, there is a natural progression from controlled and restrained to natural and actual among the levels, user, and environments. Altogether, there are numerous interdependencies among all four groups.

For the technology levels, the most basic pieces of a system are the *Sub-Components* at the lowest, testable levels. They can be viewed as very simple black boxes where there are a few simple inputs producing a few simple outputs. *Components* are a step up from *Sub-Components*. *Components* are easily identified as those constituent features that could be broken down further (into *Sub-Components*). For evaluation purposes, *Components* need to be combined with other *Components* to comprise the entire *System*. The highest technology level is the *System*. Tests at the *System* level are influenced by the most inputs (as compared to the *Component* and *Sub-Component* levels) and therefore yield a wide range of outputs. *Capabilities* are produced from *Sub-Components* and/or *Components* interacting together to produce an action. *Capabilities* can occur at both the *Component* and *System* levels depending upon the technology's makeup. Every *Tech User* group would not be able to operate the system across all of the technology levels. For example, the *Sub-Components* and *Components* are typically not system fragments that *End-Users* would see during their natural usage nor would it be practical to collect *Utility Assessments* here. Tests at these levels would be best left to the *Tech Developers* to act as the operators since they have the deepest understanding of the technology (compared to the other *Tech User* groups). Also, *Component* level (and lower) evaluations yield *Technical Performance* data as opposed to *Utility Assessment* data so the evaluation would not require *End-users* for technology feedback. Likewise, the *Tech User* pool greatly expands at the *Capability* level since this could be something that the *End-Users* could naturally use. All *Tech User* groups are viable when the technology is tested at the *System* level.

The next important piece of this relationship is that of the desired metrics including both *Technical Performance* and *Utility Assessments*. Typically, when an advanced technology or intelligent system is in its infant stages it's not ready for the *End User*. Early tests are usually conducted with *Tech Developers* as the *Tech Users* since it's likely that more issues will arise that they are better equipped to communicate about and efficiently address. Additionally, *Technical Performance* testing at these early stages can be more insightful than *Utility Assessments* to see if the technology and/or its individual components are working as they are intended. This is not to say that *Utility Assessments* are not important at the early stages. These metrics will still be useful in informing the *Technology Developers* on *Tech User* perceptions of the system, etc. As a technology matures and individual capabilities and the full system become available for testing, it becomes more practical to get *Tech User Utility Assessment* data, especially from the *End User* community. A technology is going to have an easier time being adopted by the intended *End User* community if their input is solicited during the development process.

Employing different categories of *Tech Users* within an evaluation will produce results that can range from poor to optimal performance and from improper to proper usage of the technology. It is reasonably assumed that out of all of the potential *Tech Users*, the *End Users* will have the most operational knowledge of the technologies' target usage environment, but will have the lowest understanding (and experience) using the technology. Conversely, it is reasonably assumed that those *Tech Developers* that are assigned to the *Tech Users* will have the least operational knowledge of the technologies' target usage environments, but will have the greatest (if not complete) understanding of the technology's operation. No such assumptions can be made with respect to the *Tech Users* that fall into the *Trained User* category. The individual relationships between these various personnel and their level of understanding, of both the technology and the environment, will be further explained in the following subsection.

The last significant piece of this relationship is that of the environment. Typically, emerging and/or immature technologies are evaluated in the *Lab* so that specific variables can be controlled in an effort to determine what impacts the technologies' performance and to what degree. As the technologies' further develop, they are then evaluated in less controlled environments. Tests performed in these *Simulated* environments bring the evaluators and technologists one step closer to understanding how the systems behave in more realistic environments. Ultimately, the technology is tested in the *Actual* environment once it has significantly matured and nears its final design. Of course, it is possible to test an immature technology in an environment more advanced than its development (such as the *Simulated* or *Actual*), but it will be much more difficult to pinpoint the exact cause(s) of failure when the technology falters. The opposite is true that a very mature technology may be tested in a more basic environment (such as the *Lab* or *Simulated* depending upon the stage of evolution), but it's likely that the results from these tests will be highly repeatable and therefore, not as practical (as compared to testing in a more advanced environment) to conduct after numerous test runs.

The evaluation pinnacle is to test a *System* in the *Actual* environment where it is used by *End Users*. At a minimum, *Utility Assessment* metrics could be collected to determine how well the technology aided the *End User* in accomplishing their objective(s). Depending upon the makeup of the test environment, certain *Technical Performance* metrics could be captured to assist in validating the final design. This is as close to realistic usage (if not already realistic) as possible of the technology and therefore presents the truest indicator of how the technology would perform in common practice. It is understood that intelligent, advanced, and emerging technologies must go through numerous evaluations at lesser variables within these four categories before the *System* can be tested in the *Actual* environment by the *End User*.

Relationship Among Personnel, Knowledge and Autonomy Levels

This relationship set is defined to describe the various personnel, their knowledge and their autonomy levels in an evaluation blueprint. These relationships are represented in the evaluation blueprint by a matrix presented in Table 1. For each of the personnel groups listed (*Tech-User*, *Team Member*, *Participant*), *Technical Knowledge* and *Operational Knowledge* simply means what they know while *DM Autonomy – Technical* and *DM Autonomy – Environment* refers to what they can do and are allowed to do with their respective knowledge. Specifically, knowledge and autonomy levels can range from none to low to medium to high. In some cases, an individual or group may not be given any autonomy during a test event which is represented as N/A. The varying levels of knowledge and DM autonomy specific to each personnel group are also presented in Table 1.

Table 1: Relationship - personnel, knowledge, and autonomy levels

	Tech-User	Team Member	Participant
Technical Knowledge	Low - Med - High	Low - Med - High	Low - Med - High
Operational Knowledge	Low - Med - High	Low - Med - High	Low - Med - High
DM Autonomy - Tech.	None - Low - Med - High	None - Low - Med - High	N/A
DM Autonomy - Env.	None - Low - Med - High	None - Low - Med - High	None - Low - Med - High

Each personnel group's knowledge and autonomy levels ranges from "Not Applicable (N/A)" to "High" as specified by MRED in its output evaluation blueprint. Based upon the required levels, the evaluation designer must then identify the appropriate personnel including the specific group(s) of *Tech-Users*.

"None" means that this personnel group has no knowledge in a specific area, DM authority either over the technology and/or how they behave within the environment. "Low" means that this personnel group has a small amount of knowledge or their DM autonomy is significantly limited in a specific area. "Med" (medium) means that this personnel group has an average amount of knowledge or is given some DM autonomy in a specific area. "High" means that this personnel group has expert and/or extensive knowledge or full DM autonomy in a specific area. For example, suppose the US Marines are testing an advanced combat vehicle and a Marine is employed as a *Tech User: End User*. The Marine can be categorized as having no technical knowledge of the vehicle when they are seeing it for the first time. After an hour of basic training on the vehicle, it could be reasonably stated that the Marine has "Low" technical knowledge of the system; after a week of training during some simulated situations, it could be stated that the Marine has "Medium" technical knowledge of the system; and after a month of continuous usage of the vehicle in realistic environments it could be stated that the Marine has a "High" amount of technical knowledge.

Similar statements could be made about the Marine's level of operational knowledge when they first enlist in the Marine Corps ("None" - no operational knowledge), "Low" after having finished boot camp, "Medium" after having gone through capstone training exercises, and "High" after having served a tour in combat (high operational knowledge). Typically, all *Tech Users* (no matter what sub-group they fall into) have at least "Low" technical and operational knowledge prior to the evaluation due to initial system training and/or background information on their scenario objective (to support at least a minimal amount of operational knowledge).

One dependency relationship is that of DM Autonomy on the corresponding personnel knowledge. A participant's DM autonomy level cannot exceed their knowledge level. This rule holds for the *DM Autonomy Technical & Technical Knowledge* pair and the *DM Autonomy-Environmental & Operational Knowledge* pair. For example, a *Tech Developer,* who knows the intricacies of the new technology, may be assigned as the *Tech User* for a *Capability Level Testing – Technical Performance* evaluation. This could be the result of MRED's blueprint output stating that the *Tech User's Technical Knowledge* should be high. Furthermore, MRED could further dictate that the *Tech User* should have no *DM Autonomy – Technical* ("None") in the evaluation. In effect, this becomes a scripted test. However, MRED would not output a blueprint where a *Tech User* is required to have a "High" level of *DM Autonomy – Environmental* and a lesser level ("Medium" or lower) of *Operational Knowledge*. This would enable the *Tech Users* to have authority in an area where their knowledge is limited by comparison which is not practical considering their actions and responses are likely to be inappropriate and unrepresentative (since they have not had the training or experience in the given environment).

As defined earlier, *End-Users*, *Trained Users* and *Tech Developers* (when asked to test the technology) are specific cases of *Tech Users*. For example, *End-Users* will most likely have a greater level of *Operational Knowledge* and a lower level of *Technical Knowledge*. Their *DM Autonomy* in both technical and operational categories will vary given the goal of the evaluation including the desired metrics, level of technology under test and the test environment. *Trained Users* will most likely have no ("None") to "Low" *Operational Knowledge* and *Technical Knowledge*. It is also likely that their *DM Autonomy* in both technical and operational categories will be significantly limited since their knowledge is also limited. *Tech Developers* are assumed to have no ("None") to very little ("Low") *Operational Knowledge*, but very "High" (if not expert) levels of *Technical Knowledge*. Their *DM Autonomy – Environmental* will probably be limited (due to their "Low" *Operational Knowledge)*, but their *DM Autonomy – Technical* would range from low to high according to the evaluation goal, desired metrics, technology level under test, and test environment. Of course, some exceptions may exist. For example, a former Marine may now be a *Tech Developer* on an emerging military technology.

Team Members are those individuals in the environment who support the scenario or operational objective of the *Tech Users* and/or are in a position to provide qualitative feedback on the technology's impact on the *Tech Users'* situational awareness, mission efficiency, etc. Depending upon a *Team Member's* designated function within the evaluation (whether it is to support the operational objective, solely to provide qualitative feedback on their perception of the technology, or both) they could have *Technical and Operational Knowledge* ranging from "None" to "High." Also, depending upon their assigned responsibilities within the test scenario, they could be instructed to have no direct contact with the technology (*DM Autonomy – Technical* of "None") to taking control of the technology if the primary *Tech User(s)* are having difficulty ("High" *DM Autonomy – Technical*). Additionally, their *DM Autonomy – Environmental* will range from "Low" to "High" depending upon their assigned responsibilities within the evaluation scenario. Note that *Team Members* may not be required for all evaluation types or not at all for a specific technology.

EVALUATION DESIGN EXAMPLE

To further explain the proposed evaluation framework, the authors will apply the key terms and relationships to an example evaluation design of an advanced vehicle technology. Since the initial effort of creating this evaluation framework is on numerous outputs (elements of the evaluation blueprint), only the key terms and their relationships defined in earlier sections will be discussed in depth. *Stakeholders, technology state, and available resources for testing and analysis* will not be presented with any depth.

Suppose that an automobile manufacturer has developed a new feature for its line of luxury sedans that automatically connects a person's cellular phone to the vehicle's onboard computer and automatically notifies the user of any new emails or text messages via vehicle display. The fictitious technology, being called CarComm by the manufacturer, has been designed so that the user can set the vehicle to alert them visually with a light on the display or audibly through the vehicle's speakers. Presently, CarComm is planned for implementation in the vehicle production line in two years and has already undergone several redesigns due to software issues in the on-board computer sending audible alerts. CarComm has reliably informed the driver of emails and text messages via display.

There are numerous tests with a range of variables that could be proposed to further evaluate this technology. The first step in creating the evaluation design is to determine its goals. Does the management team want to get *End User* survey feedback on CarComm's ease-of-use or does the technology design team want to measure inputs into the speakers to see if they fixed the audible alert issues? This is where *Stakeholders, Technology State* and *Available Resources for Testing and Analysis* would provide input into the evaluation framework to yield a priority of evaluation goals. To demonstrate the ability of MRED to represent evaluation blueprints suitable for different goals and stakeholders, a set of three different testing requests will be described. The MRED evaluation blueprint personnel matrix appropriate for each will be shown.

Technology Levels and Metric Types

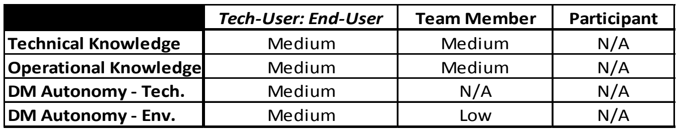
CarComm could be evaluated at multiple levels. After inputting the *Stakeholder, Technology State,* and *Available Resources for Testing and Analysis* data, suppose that MRED indicates that the top three most important evaluation goal types for execution and some relevant metrics are:

- MRED Blueprint 1. *Capability Level Testing – Technical Performance* – According to MRED, the purpose of this evaluation goal type would be to obtain quantitative metrics regarding the CarComm capability including the capture of the following:
  - Average time for phone to establish communication with CarComm
  - Average time for CarComm to download emails and texts
  - Average dB level of audible alert
  - Average time of visible alert
- MRED Blueprint 2. *System Level Testing – Utility Assessment* – MRED's output dictates that the intent of this evaluation goal type is to collect qualitative metrics regarding the CarComm capability while being operated as intended during use of the vehicle. Some metrics captured from this testing could include:
  - What did you like the most about the technology?"
  - "What did you like the least about the technology?"
  - "What would you change about the technology?"
- MRED Blueprint 3. *Component Level Testing – Technical Performance* – MRED has determined that the onboard computer component should be isolated and tested to determine if previous failures at this point have been resolved. Examples of quantitative data to be captured may include:
  - Average time for onboard computer to process data received from CarComm
  - Average time for onboard computer to output signal to speakers for audible alert
  - Average strength of the signal being sent to the speakers

Personnel

MRED configures the personnel matrices for each of the three evaluation goal types specified above. The personnel matrices for the blueprints are shown in Table 2, Table 3, and Table 4.

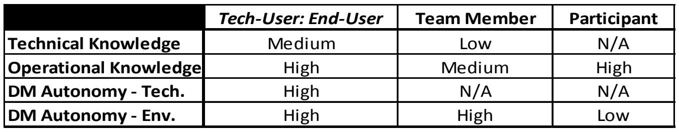
Table 2: Personnel Matrix for *Capability Level – Tech. Performance*



	Tech-User: End-User	Team Member	Participant
Technical Knowledge	Medium	Medium	N/A
Operational Knowledge	Medium	Medium	N/A
DM Autonomy - Tech.	Medium	N/A	N/A
DM Autonomy - Env.	Medium	Low	N/A

In the case of the *Capability Level Testing – Technical Performance* goal type, *End-Users* are the target group of this luxury sedan. According to MRED's blueprint, the *End-User(s)* will operate CarComm while sitting in the driver's seat. This could be followed up by additional testing where the *End-User* is sitting in the passenger's seat. *Team Members* are assigned to be the other personnel in the vehicle who are not operating CarComm where they are either sitting in the driver's seat (when the *End-User* is sitting in the passenger's seat), passenger's seat (when the *End-User* is sitting in the driver's seat) or in the backseat. MRED did not assign any *Participants* in this test blueprint.

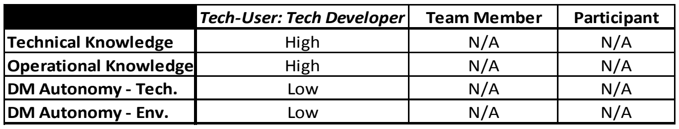
Table 3: Personnel Matrix for *System Level - Utility Assessment*



	Tech-User: End-User	Team Member	Participant
Technical Knowledge	Medium	Low	N/A
Operational Knowledge	High	Medium	High
DM Autonomy - Tech.	High	N/A	N/A
DM Autonomy - Env.	High	High	Low

In the case of *System Level Testing – Utility Assessment,* the *End-Users* and *Team Members* are in the same relative positions with similar backgrounds as the *Capability Level Testing – Technical Performance.* The most significant differences are in the levels of knowledge and autonomy each possesses. Additionally, *Participants* are present within the environment in the form of pedestrians walking around, and driving in other vehicles.

Table 4: Personnel Matrix for *Component Lvl - Tech Performance*



	Tech-User: Tech Developer	Team Member	Participant
Technical Knowledge	High	N/A	N/A
Operational Knowledge	High	N/A	N/A
DM Autonomy - Tech.	Low	N/A	N/A
DM Autonomy - Env.	Low	N/A	N/A

In the case of *Component Level Testing – Technical Performance*, *Tech Developers* are assigned by MRED as the *Tech Users* since this test is focused on operating an element of the technology that no other user base would realistically use. No *Team Members* or *Participants* are warranted for this testing.

Environment

The next evaluation element to discuss in MRED's output blueprints would be the environment(s) in which the above three evaluation goal types should be performed. One set of reasonable environment outputs given the above personnel matrices could include the following per evaluation goal type:

- *Capability Level Testing – Technical Performance* – MRED dictates that this test conducted in a *Simulated* environment which could simply be outside of the vehicle testing facility, where *Tech Users* can sit and operate the CarComm feature. In order to comply with MRED's output that no *Participants* be used, the immediate area around the test vehicle needs to be controlled to prevent unwanted interactions.
- *System Level Testing – Utility Assessment* - This environment could be a busy parking lot. There is minimal to no control over the ambient variables with vehicle and people traffic coming and going as they naturally would.
- *Component Level Testing – Technical Performance* - CarComm could be tested within an isolated *Lab* environment where the evaluation team has direct control over input into the onboard computer and can efficiently measure output.

VALIDATION OF MRED

To validate the MRED framework, the authors will analyze several previously-employed, SCORE-inspired speech-to-speech technology evaluations conducted over the past three years. This will show that the MRED framework can successfully model a pre-existing evaluation even though MRED is still in its infancy. These speech-to-speech systems are an advanced technology research and development program intent on quickly creating and fielding free-form, two-way speech-to-speech translation devices that enable personnel of different languages to communicate with one another in real-world tactical situations without the need for an interpreter [12] [24]. Since NIST has been applying the SCORE framework to design these speech-to-speech technology evaluations since 2007, multiple evaluation goal types have been defined and implemented. Metric types are defined with each evaluation goal type along with specific personnel and environments in which the evaluation was to be performed.

Technology Levels and Metric Types

After consideration of the *Stakeholders*, *Technology State*, and *Available Resources for Testing and Analysis*, the following three evaluation goal types became the predominant evaluations conducted by the NIST team in assessing the speech-to-speech technology.

- Offline Test. *Component Level Testing – Technical Performance* – Quantitative data was captured during the "Offline" evaluations to assess the performance of the technologies' three primary software components: Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS). This test was purely conducted by inputting set audio files and text utterances into each technology where the output text and audio files were analyzed. Numerous metrics were captured for each including:
    - Low-level concept transfer
    - Word error rate to assess ASR and TTS
- Lab Test. *System Level Testing – Technical Performance* – Quantitative data was collected during the "Lab" evaluations to determine how well the system could convey specific information during conversations between English and foreign language speakers in highly-controlled settings. The speakers' dialogues were controlled, but not scripted, through the use of structured scenarios which provided the speakers with the concepts they should

convey in each utterance (although the speakers were free to phrase them as they saw fit). Some of the technical performance metrics captured during the Lab evaluations included:

  - Number of questions correctly translated per 10 minutes
  - Number of attempts per question
  - Number of answers correctly translated per 10 minutes
  - Number of attempts per answer

- Field Test. *System Level Testing – Utility Assessment* – Qualitative data was collected during the "Field" evaluations to assess the utility of the technology to the target user population. Field evaluations were conducted by English and foreign language speakers holding conversations with one another where their dialogues were governed by spontaneous scenarios. This afforded them the opportunity to say whatever they wanted so long as it was in keeping with the tactical theme of the scenario. The following data was captured and analyzed in these evaluations:
  - Survey questionnaires completed by the English and foreign language speakers
  - Semi-structured interviews with the English and foreign language speakers

Personnel

For each of the three evaluations defined above, personnel were identified for each and are listed in MRED personnel matrices in Table 5, Table 6, and Table 7 below.

Table 5: Personnel Matrix for Offline Evaluation

	Tech-User: Tech Developer	**Team Member**	**Participant**
Technical Knowledge	High	N/A	N/A
Operational Knowledge	High	N/A	N/A
DM Autonomy - Tech.	None	N/A	N/A
DM Autonomy - Env.	N/A	N/A	N/A

The Offline evaluation required *Tech Developers* to operate the technology since this involved feeding audio and text data directly into the system which would not be practical by any other *Tech User* group. For the sake of autonomy, this evaluation can be considered scripted since the *Tech Developers* were instructed to input pre-defined (by the evaluation team) files into their systems.

Table 6: Personnel Matrix for Lab Evaluation

	Tech-User: End-User	**Team Member**	**Participant**
Technical Knowledge	Low	N/A	Low
Operational Knowledge	High	N/A	High
DM Autonomy - Tech.	Low	N/A	N/A
DM Autonomy - Env.	Low	N/A	Low

The Lab evaluations called for *End-Users*, Soldiers and Marines with experience interacting with foreign language personnel, to operate the technology after receiving several hours of basic training on the system. Given the nature of the lab and their limited training, their *DM Autonomy – Technical* was considerably restricted and they had little freedom in their *DM Autonomy – Environmental* given the nature of the structured scenarios that supported their conversations. The *Participants* in this evaluation were the foreign language speakers who responded to the English speakers' questions. These participants were given no control over the technology and were also restricted in their dialogues by the structured scenarios.

Table 7: Personnel Matrix for Field Evaluation

	Tech-User: End-User	**Team Member**	**Participant**
Technical Knowledge	Medium	N/A	Low
Operational Knowledge	High	N/A	High
DM Autonomy - Tech.	Medium	N/A	N/A
DM Autonomy - Env.	Medium	N/A	Medium

The Field evaluations also called for the Soldiers and Marines (*End-Users*) to operate the technology to communicate with foreign language speaking *Participants.* Since these evaluations always occurred after the Lab evaluations, the *End-Users* had greater technical knowledge (due to their increased experience) in the Field as compared to the Lab. Additionally, since these evaluations were intended to be under more realistic conditions, the *End-Users* had a more *DM Autonomy – Technical* with the speech-to-speech systems. Both the *End Users* and *Participants* had more *DM Autonomy – Environmental* since their conversations were governed by spontaneous scenarios where they could speak about whatever they wished so long as it was within specific tactically-relevant domains.

Environment

Each of the three main speech-to-speech technology evaluation types were set in unique environments.

- Offline – This evaluation took place in a *Lab* environment, specifically in a conference room where audio and text data was input into the systems via USB thumb drive. The output data was also collected on the same thumb drive and taken for analysis by the *Evaluators*.
- Lab – This evaluation also took place in a *Lab* environment in the form of individual conference rooms where each room supported a single technology with an assigned English speaker and foreign language speaker. The only other personnel allowed in the room were *Evaluators* who collected pertinent data. Each room was isolated from the outside so that no ambient noise could disturb the test event.
- Field – This evaluation was conducted in a *Simulated* environment which took the form of a field and secluded roadway on NIST grounds. Specifically, this environment afforded the opportunity for the *End-Users* and *Participants* to move about more freely with the technology in areas where outside noise was present, but maintained at a distance.

Notice that the SCORE-prescribed evaluations possess the evaluation blueprint characteristics that would be output from applying the MRED framework. These characteristics were chosen by the evaluation team based upon their extensive experience in evaluation design. Ultimately, the development of MRED and its application to those technologies requiring evaluation, will lead to the automatic generation of specific evaluation blueprints no matter the experience of the design team.

FURTHER APPLICATION OF MRED

Additional evaluation design projects are currently being examined as potential MRED applications. One such project is the assessment and evaluation of multiple pedestrian tracking algorithms whose test design and implementation is conducted jointly by NIST and members of the Army Research Laboratory's (ARL) Collaborative Technology Alliance (CTA) [3]. Specifically, the ARL CTA project is evaluating algorithms produced from numerous companies and organizations which use Laser Detection and Ranging (LADAR) and video sensor data taken from a moving platform. The evaluation team employs these sensors on a vehicle, where the vehicle moves through a test environment and the vehicle-mounted sensors collect and feed data to on-board detection and tracking algorithms.

To date, the ARL CTA/NIST team has collaboratively planned and implemented several evaluations from 2007 through 2010. In an effort to further expand the test capabilities of this project, this work will be discussed in terms of the initial MRED framework design. The ARL CTA test design will be correlated to MRED's technology and metric levels, personnel and the environment.

Technology Levels and Metric Types

Based upon the current level of maturity of the technology, the ARL CTA is very focused on isolating the pedestrian detection and tracking algorithms in a manner that will yield quantitative technical performance metrics. NIST's involvement with the program has centered on conducting field exercises in the category of *Capability Level Testing – Technical Performance*. These field exercises capture the technical data required to assess the performance of the CTA teams' algorithm. They also provide data to support future algorithm development and produce performance analyses that are based on the collected data to aid obstacle avoidance planning [3]. These exercises are conducted by having a sensor-laden vehicle drive in a pedestrian- and obstacle-filled environment where experimental algorithm output could be collected and measured against evaluation team-captured ground truth.

Based upon the captured experimental and ground truth data, numerous performance metrics were applied including:

- Detection – This includes noting whether humans are correctly identified, misclassified as non-human obstacles, incorrectly mistaken for unknown course features or not detected at all.
- Moving vs. Static Entities
- Entity Classification for Non-humans – Are objects classified as barrels, cones, crates, etc.

Personnel

For the evaluation type defined above, personnel were identified for each and are shown in the MRED personnel matrix in Table 8 below.

Table 8: Personnel Matrix for ARL CTA Evaluation

	Tech-User: Trained User	Team Member	Participant
Technical Knowledge	Low	N/A	N/A
Operational Knowledge	Low	N/A	Low
DM Autonomy - Tech.	N/A	N/A	N/A
DM Autonomy - Env.	Low	N/A	Low

This evaluation required a *Trained User* to engage and disengage the technology during the test runs. Since this capability will ultimately be integrated into a greater system and is still relatively immature, it is too soon to identify the intended exact user group. Algorithms from multiple organizations were operating at once and engaged by the same *Tech User*, yielding a more objective approach than selecting specific *Tech Developers* to operate the technology.

The evaluation participants were those individuals that acted as pedestrians, also known as "walkers." They were given a specific path within the environment that they walked during the tests. Practice runs were conducted so that the walkers could determine their pace, better enabling them to complete their path in a prescribed amount of time.

Environment

These tests were conducted in *Simulated* environments that included some Military Operations in Urban Terrain (MOUT) features. The evaluation team controlled the environment during the testing process allowing it to get very detailed ground-truth data including walking paths, obstacle locations, and vehicle path. The evaluation team also closed down the area to non-evaluation personnel for both safety and test quality purposes.

Figure 3 presents an overhead image of the test environment with a 10-meter grid. The blue lines represent the ground truth of the vehicle path and walker paths while the other colored lines depict the detection results of fusing all algorithms and the sensor data.

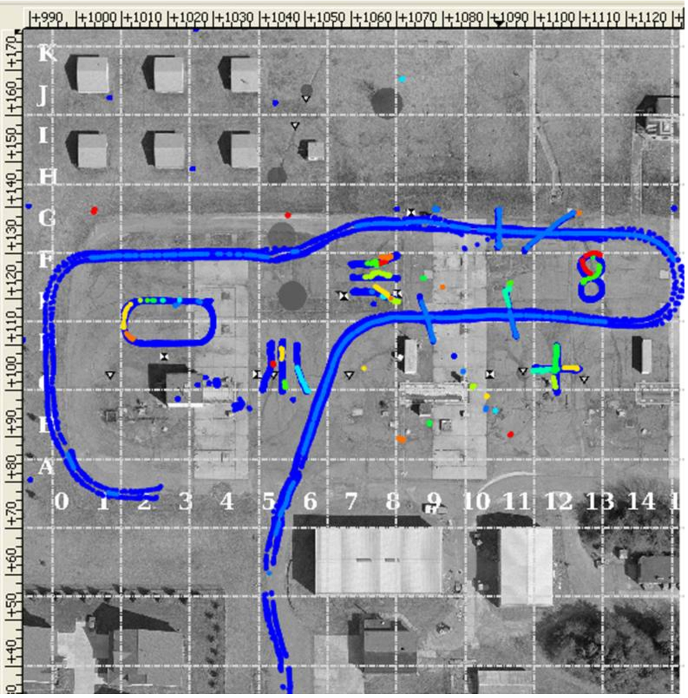


Figure 3 – Overhead image of test environment

CONCLUSION

The foundation of the Multi-Relationship Evaluation Design (MRED) framework has been established upon the successes of the SCORE framework. MRED's evaluation blueprint elements are built upon the relationships between technology levels, metric types, personnel, and test environments. This initial work has also defined a representation model for the evaluation personnel's knowledge levels and corresponding decision-making autonomies. These MRED blueprint element models were successfully demonstrated within the vehicle technology. Finally, the MRED personnel matrix model's adequacy and adaptability is demonstrated in the example along with showing their applicability in a previously-defined, SCORE-inspired evaluation design of an advanced technology.

In future work, MRED will be expanded with the definition of additional evaluation blueprint elements including environmental factors, data collection methods, evaluation scenarios, and data analysis methods. Further work will be conducted to discuss MRED's inputs, their associated uncertainties and how MRED ultimately processes this information to output a comprehensive evaluation blueprint including the recognition of uncertain evaluation elements.

ACKNOWLEDGEMENTS

The authors acknowledge and appreciate the support of the ARL CTA program.

REFERENCES

[1] Albus, J.S., Barbera, A.J., Scott, H.A., Balakirsky, S.B., 2006, “Collaborative Tactical Behaviors for Autonomous Ground and Air Vehicles,” *Proc. of the Unmanned Ground Vehicle Technology VII – SPIE Conference*, **5804**, pp. 244-254.

[2] Balakirsky, S.B. and Madhavan, R., 2009, “Advancing Manufacturing Research Through Competitions,” *Proceedings of the SPIE Defense, Security, and Sensing Conference*.

[3] Bodt, B., Camden, R., Scott, H., Jacoff, A.S., Hong, T., Chang, T., Norcross, R., Downs, T., Virts, A., 2009, “Performance Measurements for Evaluating Static and Dynamic Multiple Human Detection and Tracking Systems in Unstructured Environments,” *Proc. of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Workshop*.

[4] Bostelman, R.V., Hong, T.H., Madhavan, R., Chang, T.Y., and Scott, H.A., 2006, “Performance Analysis of Unmanned Vehicle Positioning and Obstacle Mapping,” *Proc. of the Unmanned Systems Technology VIII - SPIE Conference*, **6230**(2).

[5] Calisi, D., Iocchi, L., and Nardi, D., 2008, “A Unified Benchmark Framework for Autonomous Mobile Robots and Vehicles Motion Algorithms (MoVeMA benchmarks),” *In Workshop on Experimental Methodology and Benchmarking in Robotics Research (RSS 2008)*.

[6] Conley, S.A., 2009, “Test and Evaluation Strategies for Network-Enabled Systems,” International Test and Evaluation Association (ITEA) Journal, **30,** pp. 111-116.

[7] Jacoff, A.S., Downs, A.J., Virts, A.M., and Messina, E., 2008, “Stepfield Pallets: Repeatble Terrain for Evaluating Robot Mobility,” *Proc. of the 2008 Performance Metrics for Intelligent Systems (PerMIS) Workshop*, pp. 29-34.

[8] Jacoff, A.S., and Messina, E., 2007, “Urban Search and Rescue Robot Performance Standards: Progress Updated,” *Proc. of the Unmanned Systems Technology IX – SPIE Conference*, G.R. Gerhart et al., eds., **6561**, pp. 65611L.

[9] Messina, E., 2009, “Robots to the Rescue,” Crisis Response Journal, **5**(3), pp. 42-43.

[10] Messina, E. and Jacoff, A.S., 2007, “Measuring the Performance of Urban Search and Rescue Robots,” *IEEE Conference on Homeland Security Technologies*.

[11] Schlenoff, C.I., Steves, M.P., Weiss, B.A., Shneier, M.O., and Virts, A.M., 2007, “Applying SCORE to Field-Based Performance Evaluations of Soldier-Worn Sensor Technologies,” Journal of Field Robotics – Special Issue on Quantitative Performance Evaluation of Robotic and Intelligent Systems, **24**, pp. 671-698.

[12] Schlenoff, C.I., Weiss, B.A., Steves, M.P., Sanders, G., Proctor, F., and Virts, A.M., 2009, “Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies,” *Proc. of the Performance Metrics for Intelligent Systems (PerMIS) Workshop*.

[13] Schlenoff, C.I., Weiss, B.A., Steves, M.P., Virts, A.M., and Shneier, M.O., 2006, “Overview of the First Advanced

Technology Evaluations for ASSIST," *Proc. of the Performance Metrics for Intelligent Systems (PerMIS) Workshop.*

[14] Scholtz, J.C., Antonishek, B., and Young, J.D., 2005, "A Field Study of Two Techniques for Situation Awareness for Robot Navigation in Urban Search and Rescue," *Proc. Of the IEEE Ro-Man Conference.*

[15] Scholtz, J.C., Theofanos, M.F., and Antonishek, B., 2006, "Development of a Test Bed for Evaluating Human-Robot Performance for Explosive Ordnance Disposal Robots," *Proc. Of the 1st Annual Conference on Human-Robot Interaction*.

[16] Scholtz, J.C., Antonishek, B., and Young, J.D., 2004, "Evaluation of Human-Robot Interaction in the NIST Reference Search and Rescue Test Arenas," *Proc. of the 2004 Performance Metrics for Intelligent Systems (PerMIS) Workshop.*

[17] Scrapper, C.J., Madhavan, R., Balakirsky, S.B., 2008, "Performance Analysis for Stable Mobile Robot Navigation Solutions," *Proc. of the Unmanned Systems Technology X – SPIE Conference*, **6962**(6), pp. 1-12.

[18] Steves, M.P., 2007, "Utility Assessments of Soldier-Worn Sensor Systems for ASSIST," *Proc. of the 2006 Performance Metrics for Intelligent Systems (PerMIS) Workshop*.

[19] Sukhatme, G.S. and Bekey, G.A., 1995, "An Evaluation Methodology for Autonomous Mobile Robots for Planetary Exploration," *Proc. of the First ECPD International Conference on Advanced Robotics and Intelligent Automation*, pp. 558-563.

[20] Thompson, M., 2008, "Testing the Intelligence of Unmanned Autonomous Systems," International Test and Evaluation Association (ITEA) Journal, **29**, pp. 380-387.

[21] Weiss, B.A., Menzel, M., 2009, "Development of Domain-Specific Scenarios for Training and Evaluation of Two-Way, Free Form, Spoken Language Translation Devices," *Proc. of the 2009 International Test and Evaluation Association (ITEA) Symposium*.

[22] Weiss, B.A. and Schlenoff, C.S., 2008, "Evolution of the SCORE Framework to Enhance Field-Based Performance Evaluations of Emerging Technologies," *Proc. of the 2008 Performance Metrics for Intelligent Systems (PerMIS) Workshop*, pp. 1-8.

[23] Weiss, B.A., and Schlenoff, C.I., 2009, "The Impact of Scenario Development on the Performance of Speech Translation Systems Prescribed by the SCORE Framework," *Proc. of the Performance Metrics for Intelligent Systems (PerMIS) Workshop*.

[24] Weiss, B.A., Schlenoff, C.I., Sanders, G.A., Steves, M.P., Condon, S., Phillips, J., and Parvaz, D., 2008, "Performance Evaluation of Speech Translation Systems," *Proc. of the 6th edition of the Language Resources and Evaluation Conference*.

[25] Weiss, B.A., Schlenoff, C.I., Shneier, M.O. , and Virts, A.M., 2006, "Technology Evaluations and Performance Metrics for Soldier-Worn Sensors for ASSIST," *Proc. of the 2006 Performance Metrics for Intelligent Systems (PerMIS)Workshop*.