The Multi-Relationship Evaluation Design Framework: Creating Evaluation Blueprints to Assess Advanced and Intelligent Technologies

Brian A. Weiss
National Institute of Standards and Technology
100 Bureau Drive, MS 8230
Gaithersburg, Maryland 20899
1-301-975-4373
brian.weiss@nist.gov

Linda C. Schmidt
University of Maryland
0162 Glenn L. Martin Hall, Building 088
College Park, Maryland 20742
1-301-405-0417
lschmidt@umd.edu

ABSTRACT
Technological evolutions are constantly occurring across advanced and intelligent systems across a range of fields including those within the military, law enforcement, automobile, and manufacturing industries. Testing the performance of these technologies is critical to (1) update the system designers of areas for improvement, (2) solicit end-user feedback during formative tests so that modifications can be made in future revisions, and (3) validate the extent of a technology's capabilities so that both sponsors, purchasers and end-users know exactly what they are receiving. Evaluation events can be minimally designed to include a few basic tests of key technology capabilities or they can evolve into extensive test events that emphasize multiple components and capabilities along with the complete system, itself. Tests of advanced and intelligent systems typically assume the latter and can occur frequently based upon system complexity. Numerous evaluation design frameworks have been produced to create test designs to appropriately assess the performance of intelligent systems. While most of these frameworks allow broad evaluation plans to be created, each framework has been focused to address specific project and/or technological needs and therefore has bounded applicability. This paper presents and expands upon the current development of the Multi-Relationship Evaluation Design (MRED) framework. Development of MRED is motivated by the desire to automatically create an evaluation framework capable of producing detailed evaluation blueprints while receiving uncertain input information. The authors will build upon their previous work in developing MRED through an initial discussion of key evaluation design elements. Additionally, the authors will elaborate upon their previously-defined relationships among evaluation personnel to define evaluation structural components pertaining to the evaluation scenarios, test environment, and data collection methods. These terms and their relationships will be demonstrated in an example evaluation design of an emerging technology.

PerMIS'10, September 28-30, 2010, Baltimore, MD, USA.

Categories and Subject Descriptors
B.8.0 [**Performance of Systems**]: m*easurement techniques, modeling techniques, performance attributes.*

General Terms
Measurement, Performance, Design, Experimentation, Verification.

Keywords
MRED, SCORE, performance evaluation, model, framework

1. INTRODUCTION
Advanced technologies and intelligent systems are emerging across a range of domains including those within the military, law enforcement, automobile, manufacturing and oil industries. An example of a technology are the remotely operated underwater vehicles (ROVs) currently being used to support the Gulf of Mexico oil spill [5]. One commonality of these systems is the human robot interface (HRI) or human computer interaction (HCI) component [11] [12]. Evaluating the performance of these intelligent systems is of paramount importance to (1) inform the technology designers of shortcomings, (2) solicit end-user feedback, and (3) validate the technology's final capabilities. The former occurs in formative evaluations so that modifications can be made in upcoming design iterations; the latter occurs in summative evaluations so that buyers and technology users know exactly what they are getting. These HRI and HCI technologies still feature human-in-the-loop operation. The user's involvement with the technology can range from having full control over all system functions to simply monitoring the system's behavior and can include dynamically varying the levels of control between these two limits.

Both formative and summative evaluations can be minimally structured to include several basic tests of key system capabilities or they can take the form of comprehensive test events that focus on multiple sub-system components and capabilities [8]. Evaluation events of advanced and intelligent systems usually focus on these multiple levels. These tests can justifiably occur more frequently based upon their inherent system complexity.

Extensive evaluations of emerging and intelligent technologies have occurred in numerous domains. Examples include the evaluations of autonomous ground vehicles along with several constituent components (i.e., intelligent control architectures, automated positioning and mapping technologies, obstacle and pedestrian tracking systems) [1] [2] [13]. Likewise, considerable

resources have been exerted to test the advanced technologies of Urban Search and Rescue (US&R) and bomb disposal robotic systems. To date, a widespread range of tests have been designed, fabricated, implemented, and iterated to test US&R and bomb disposal robots across a collection of operational situations [6] [7]. The tests designed to evaluate these technologies range from specific test methods aimed at assessing individual system capabilities to scenarios targeted at testing the entire system.

Assessing the performance of advanced and intelligent systems has motivated research into creating methods and frameworks to design evaluation plans. Many of the frameworks developed have been sufficient to evaluate given technologies and accomplish program-specific objectives. To date, no individual framework has been recognized as being suitable to attain both quantitative and qualitative performance across a range of virtual and physical systems including those with both human-controlled and autonomous functions.

The National Institute of Standards and Technology (NIST) has created the System, Component, and Operationally-Relevant Evaluation (SCORE) framework to evaluate emerging and intelligent systems at various levels [8]. SCORE has been effectively applied to fifteen evaluations across several technologies [9] [10] [17] [19]. SCORE enabled tests have yielded extensive quantitative and qualitative data that has proven valuable to the technology developers, evaluation designers, potential end-users, and funding sponsors.

Weiss, a co-developer of SCORE, has drawn upon that success to introduce a new evaluation framework that will automatically generate evaluation blueprints (test plans). This new evaluation plan design tool is known as the Multi-Relationship Evaluation Design (MRED) framework. MRED's ultimate objective is to take inputs from three specific groups, each complete with their own uncertainties, and output an evaluation blueprint that specifies all characteristics of the tests [18]. MRED's evaluation blueprint is defined as a detailed technology evaluation 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 present the following: the author's initial development of the MRED framework. The discussion will include those elements leveraged from SCORE and further expansion of the MRED framework. MRED will be validated by applying it to an evaluation design to test an emerging technology. Finally, strategies to further develop and augment MRED will be stated.

2. OTHER FRAMEWORKS

Development of MRED is motivated by the desire to create an evaluation framework capable of producing detailed evaluation blueprints while factoring in numerous uncertainties. There are currently many test development systems that are used to evaluate complex advanced and intelligent systems. For instance, an evaluation framework was produced to test mobile robots for planetary exploration across relevant terrains [15]. However, evaluations did not consider HRI factors. Likewise, Calisi et al. [3] have devised an evaluation framework to specifically assess intelligent algorithms. Its success has been well-documented in capturing technical performance in the virtual world, yet it has not been employed to capture feedback from human users or evaluate physical systems.

The SCORE framework was created to evaluate technologies at the component level, capability level, and system level across numerous environments from highly-controlled laboratory settings to real use-case domains [16]. To date, SCORE has been successful in allowing evaluation designers to recognize the most practical blueprints for evaluating a range of intelligent technologies. MRED not only leverages some of the successes of the SCORE framework in its own design, but it also introduces several innovative features. They include (1) MRED's ability to identify relationships and interdependencies among many evaluation elements and (2) an ability to address the uncertainties from the various evaluation inputs including how they impact the blueprints.

Due to SCORE's success in identifying evaluation designs for testing speech-to-speech translation technologies, advanced soldier-worn sensor systems, along with mapping and navigation algorithms, MRED will adapt the SCORE framework's prescribed evaluation goal types [9] [10] [14] [17] [19]. These will be discussed in subsequent sections as the MRED framework is presented.

3. MRED MODEL

The Multi-Relationship Evaluation Design (MRED) model is introduced by presenting the significant design inputs and the features of the output "evaluation blueprint." These inputs and outputs are shown in Figure 1.

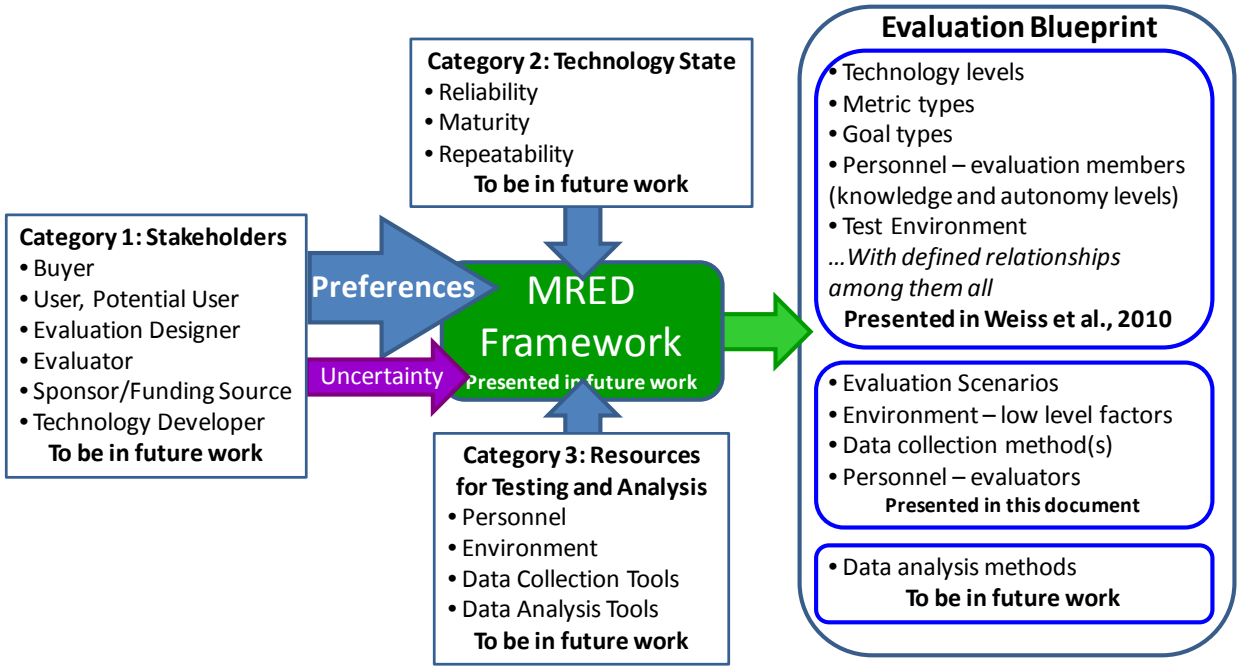

Figure 1. Input (Categories 1 to 3) and Output (Evaluation Elements) of the MRED Model

Development of the MRED model began with identifying elements of the evaluation blueprint [18]. The unique parts of this paper are the expansion and elaboration of three evaluation blueprint elements. They are *Explicit Environmental Factors*, *Data Collection Methods*, *Evaluation Scenarios*, and *Personnel - evaluators* (shown in Figure 1). Additionally, previously-identified parts of the evaluation blueprint will be presented section 3.3 as shown in Figure 1 [18].

3.1 Example MRED Application

These MRED pieces will be applied to a technology that is currently being tested by NIST personnel as each of the critical input categories and output blueprint criteria are discussed. The selected 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) [2]. Specifically, the CTA/NIST testing is focused on evaluating algorithms produced from numerous companies and organizations which use Laser Detection and Ranging (LADAR) and video sensor data taken from a moving test vehicle. This vehicle travels through the test environment and the vehicle-mounted sensors collect and feed data to the on-board detection and tracking algorithms.

From 2007 to the present day, the CTA/NIST team has jointly planned and implemented several evaluations. To expand the evaluation capabilities of the MRED, the ARL work will be discussed using the terms of the initial MRED framework design.

3.2 Input Categories

The MRED framework identifies three critical input groups that provide data into the planner. Each group will be briefly described in the following subsections. These categories will be further elaborated upon including their relationships and sources of uncertainty in future efforts.

3.2.1 Category 1 – Stakeholders

Test stakeholders are classified into six categories or parties interested in a technology's evaluation. Stakeholders could have an impact over the design of a technology evaluation. Members of these categories have their own motivation in the test plan and interests in the results of a technology's performance. Their individual motivations will reflect personal uncertainties based upon their changing preferences. An example of uncertainty within stakeholder preferences could be the sponsor's expectation of what system capabilities are crucial for testing. Based upon uncertain and/or changing information, directives from their superiors, etc, the sponsor's preference of what capabilities should be tested could be moving a target. The six personnel categories are summarized in Table 1.

Table 1. Personnel with a Stake in a Technology Evaluation

STAKEHOLDER GROUPS	WHO THEY ARE…
Buyers	Stakeholder purchasing the technology
Users, Potential Users	Stakeholder that will be, or are already using the technology
Evaluation Designers	Stakeholder creating the test plans by determining MRED inputs
Evaluators	Stakeholder implementing the evaluation test plans
Sponsors/Funding Sources	Stakeholder paying for the technology development and/or evaluation
Technology Developers	Stakeholder designing and building the technology

There may be some overlap among the stakeholders which occurs on a technology-by-technology basis. Figure 2 presents the potential relationships among the stakeholders.

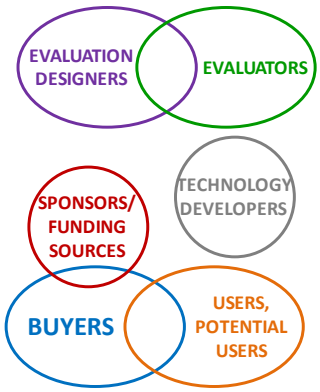

Figure 2. Stakeholder Relationships

3.2.2 Category 2 – Technology State Factors

This class or category comprises the factors that influence the technology's state at the time of its test. These factors include:

- *Reliability* – This term defines the technology's ability to be evaluated under certain conditions and/or use specific functionalities. Reliability is important because it determines if the technology is robust enough to undergo specific tests and/or if its current level of reliability limits the tests it can perform.
- *Maturity* – This term describes the technology's state or quality of being fully developed. This factor is critical because it states the degree to which the technology is equipped with all of its intended functionalities. Only a subset of expected features may be operational at the time of testing.
- *Repeatability* – This term refers to the technology's ability to yield the same or comparable results as determined from previous test(s). Repeatability is a significant factor that notes the degree to which the technology has undergone previous testing. The output test data may be used to iterate upon the design along with provide baseline data for future testing.

Understanding each of these factors will provide knowledge as to to the high-level intent of the test. The evaluation will either output formative data (intended to inform on a technology's design while it's still in development and not fully mature) or summative data (intended to validate the final design of a technology) [14].

3.2.3 Category 3 – Resources for Testing and Analysis

This last input group is composed of various types of material, personnel and technology to be included in the evaluation exercise. Resource availability (or lack thereof) and resource limitations can have a tremendous influence on the final evaluation design.

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

3.3 Previously-defined Blueprint Outputs

Previously-defined key terms in MRED's evaluation blueprint are presented here prior to introducing the new blueprint concepts. These existing terms include technology levels and metric types, which are also combined to form goal types. Additionally, evaluation personnel and environments are discussed. These previously-defined outputs were applied to the ARL CTA test case in earlier work [18]. An example will be briefly presented in the following subsections to better enable understanding of how the newly-defined outputs are applied.

3.3.1 Technology Levels

A system (often called a "technology") is made up of constituent components and they can be evaluated at multiple levels. There are several terms related to technology levels 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.

3.3.2 Metric Types

Evaluations are capable of capturing two distinct types of metrics. In defining the two metric types, it is essential to define metrics and measures in the context of the MRED model.

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

Specifically, the two metric types are:

- *Technical Performance* – Metrics related to quantitative factors (such as accuracy, precision, time, distance, etc). These metrics may be needed by the program *Sponsor* to get a status of the technology's current performance, update the *Technology Developers* on their design, etc.
- *Utility Assessments* – Metrics related to the qualitative factors that judge the condition or status of being useful and usable to the target user population. Like *Technical Performance*, these metrics may be of value to any of the stakeholders.

3.3.3 Goal Types

Goal types, extracted from the SCORE framework, are combinations of technology levels and desired metrics [9] [16]. There are five goal types employed in the MRED framework (shown in Table 2).

Table 2. Goal Types Employed by the MRED Framework

TECHNOLOGY LEVEL	METRIC TYPE	DESCRIPTION
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.

Each of these goal types requires different blueprint components and characteristics within an evaluation. Specific goal and metric types make it possible to design evaluations to collection the necessary quantitative and/or qualitative data

Due to the current level of technology maturity, the ARL CTA is currently isolating the pedestrian detection and tracking algorithms to yield technical performance metrics. Based upon this knowledge, NIST's involvement in the program has focused on conducting exercises in the *Goal Type* of *Capability Level Testing – Technical Performance*. Further discussion of *Metrics* can be found in [2] [18].

3.3.4 Personnel – Evaluation Members

Various individuals and groups are required to perform an effective evaluation. They are classified into two categories: primary (direct interaction) technology users and secondary (indirect interaction or evaluation support). The primary technology users are defined as *Tech User*. These individuals directly interact with the technology during the evaluation. They 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* that have been extensively defined in previous efforts [18]. *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*, yet are not *End-Users*.
- *Tech User: Tech Developer* – Members of the research and development organization that developed the technology under evaluation.

The secondary personnel feature those that indirectly interact with the technology during the evaluation and fall into three categories:

- *Team Member* – Individuals that work with *Tech Users* during the evaluation as they would to realistically support the use-case scenario in which the technology is immersed. Team *Members* may or may not be in a position to indirectly or directly interact with the technology during the evaluation, but they are often in a position to observe a *Tech User's* interactions with the system.
- *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.
- *Evaluator* – Personnel on the evaluation team present within the *Test Environment* that task the *Participants* and/or captures data, but do not interact with the technology. Depending upon the test, the *Evaluator* may interact with the *Tech User* to capture data.

As discussed in previous efforts [18] each of these *Personnel Groups* (excluding evaluators) has varying *Knowledge Levels* about the technology (*Technical Knowledge*) and the testing and/or use-case environment (*Operational Knowledge*). Additionally, *Autonomy Levels* are identified for these personnel groups where each has varying decision-making authority regarding the technology they were using (*DM Autonomy – Technical*) and their interactions with other personnel and the environment (*DM Autonomy – Environmental*).

Presently, the ARL CTA test effort calls for a *Tech User: Trained User* to operate the technology during the test runs. It should be noted that since the capability being tested will ultimately be incorporated into a larger system, it is premature to recognize the intended user group. Prior tests have not featured any *Team Members*, yet include numerous *Participants*. These *Participants* play the role of "walkers" where they are assigned to walk specific paths during the test. *Knowledge* and *Decision-making Autonomy* levels for these *Evaluation Members* can be found in [18].

3.3.5 Test Environments

The setting in which the evaluation occurs can have a significant effect on the data since the environment can influence the behavior of the personnel and can limit which levels of a technology can be tested. MRED defines three distinct environments that are:

- *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.

Significant relationships exist between Technology Levels, Metric Types, Tech Users and *Test Environments* which were highlighted in previous efforts [18]. Likewise, extensive relationships have been noted between *Personnel*, *Knowledge Levels*, and *Autonomy Levels* have been documented in this same effort.

The ARL CTA testing is currently being conducted in *Simulated* environments that include some Military Operations in Urban Terrain (MOUT) characteristics. This type of venue enabled the evaluation team to control many key parameters and variables within the *Test Environment*. It also allowed them collect extensive ground truth necessary for calculating several of the quantitative metrics.

3.4 Latest Key Blueprint Outputs

This work defines and further illustrates several new output elements from MRED. Specifically, they are *Evaluation Scenarios*, *Explicit Environmental Factors*, and *Data Collection Methods*. Although introduced in previous work, *Evaluators* (Personnel) will be discussed again in the context of their responsibilities and interactions with the three latest output categories. Relationships that link these blueprint components will be also be explored.

3.4.1 Evaluation Scenarios

The *Evaluation Scenarios* govern exactly what the technology will encounter and the challenges it will have to meet within the identified *Test Environments*. Three unique types of *Evaluation Scenarios* are identified below where each is unique in the relationships they have with *Tech User: Knowledge Levels*, *Tech User: Decision-Making Autonomy*, and *Environment – High Level Venues*. The three *Evaluation Scenario* types are:

- *Technology-based* – Evaluation scenarios in this category feature specific instructions to the user in how they should use the technology within the testing environment.
- *Task/Activity-based* – Evaluation scenarios in this category state the user complete a specific task within the environment where they may use the technology as they see fit.
- *Environment-based* – Evaluation scenarios in this category enable the user to perform the relevant activities within the environment based upon an advanced Operational Knowledge.

Typically, *Technology-based Evaluation Scenarios* occur in the *Lab* or *Simulated* environments where the evaluation team can determine the exact test parameters and control the various test variables. Likewise, *Task/Activity-based Evaluation Scenarios* can occur across any of the three (*Lab*, *Simulated*, *Actual*) environments where the evaluation team still has specific measures of control of both the test parameters and variables. The *Environment-based Evaluation Scenarios* can only occur in the *Simulated* and *Actual* environments where the evaluation team has no control over test parameters and variables. The specific relationships among the *Evaluation Scenarios* and the *Tech User's Knowledge Levels* and *Decision-making Autonomy* are shown in Table 3. Refer to [18] for a detailed presentation of *Knowledge and Decision-Making Autonomy Levels*.

The *Evaluation Scenarios* designed for the ARL CTA testing can be classified as *Technology-based* (see Table 4). Specifically, the *Tech User* is restricted in how they can interact with the technology. They are only allowed to engage the technology at the beginning of a run and disengage it at the run's conclusion. It is clear that the *Evaluation Scenarios* are neither *Task/Activity-based* nor *Environment-based* since the *Tech User* has no freedom to interact with the environment or has to complete a specific mission with the technology during the testing.

Since the *Evaluation Scenarios* are *Technology-based, meaning* the *Tech User* is fully-constrained as to when they can use the system, the *Tech User* then has a *DM Autonomy-Technical* value of "None." Likewise, since *Evaluators* drove the test vehicle around the site, the *Tech User* has a *DM Autonomy-Environmental* of "None." The *Tech User* has no other responsibilities and more specifically, is a *Trained User*, with some working knowledge of the technology thereby specifying the *Technical Knowledge Level* of "Medium." Multiple tracking algorithms from different organizations were evaluated simultaneously. So, it was not practical or proper to have a *Tech Developer* engage the technologies. Since the *Tech User* had no control over their activities within environment nor was it a place they had prior experience, their *Operational Knowledge* can be defined as "Low."

3.4.2 Explicit Environmental Factors

The *Explicit Environmental Factors* are significant characteristics within the environment that impact the technology and therefore, influence the outcome of the evaluation. These factors pertain to the overall physical space which is composed of *Participants* (constituent actors), structures along with any integrated props and artifacts. These factors are broken down into two characteristics, *Feature Density* and *Feature Complexity*. Together, these two elements determine the *Overall Complexity* of the environment.

- *Feature Density* – Refers to the number of features within the *Test Environment* given the size of the test area. The greater the *Feature Density*, the more challenging it is for a technology to effectively and efficiently interact with, identify objects/events/activities, operate within, etc. The *Test Environment*. *Feature Density* of a testing environment can be characterized as "Low," "Medium," and "High" referring to the level within the testing environment.

Table 3. Relationship Among the Evaluation Scenarios, Test Environments, Knowledge and Decision-making Autonomy

EVALUATION SCENARIOS	TEST ENVIRONMENT(S)	TECH USER'S KNOWLEDGE LEVEL		TECH USER'S DECISION-MAKING AUTONOMY	
		TECHNICAL	*OPERATIONAL*	*TECHNICAL*	*ENVIRONMENTAL*
Technology-based	Lab, Simulated	MED - HIGH	LOW - MED - HIGH	NONE - LOW	NONE - LOW
Task/Activity-based	Lab, Simulated, Actual	LOW - MED - HIGH	LOW - MED - HIGH	LOW - MED - HIGH	LOW - MED - HIGH
Environment-based	Simulated, Actual	MED - HIGH	MED - HIGH	MED - HIGH	MED - HIGH

- *Feature Complexity* – Refers to the intricacy of various features within the environment. For example, a baseball (sphere) has a lower *Feature Complexity* as compared to a car. Similar to *Feature Density*, the greater the *Feature Complexity*, the more difficult it is for the technology to accurately and appropriate operate and be beneficial to the *Tech User(s)*. As with *Feature Density, Feature Complexity* can also be characterized as "Low," "Medium," and "High" referring to the level within the testing environment.
- *Overall Complexity* – This factor refers to the global combination of *Feature Density* and *Feature Complexity* within the testing environment. *Overall Complexity* can range from "Low," "Low/Medium," "Medium," "Medium/High," and "High" since it integrates both density and complexity.

Table 4. ARL CTA Test Evaluation Scenario, Environment, and Tech User Parameters

EVALUATION SCENARIOS	TEST ENV.	KNOWLEDGE		DM AUTONOMY	
		TECH.	*OP.*	*TECH.*	*ENV.*
Technology-based	Simulated	MED	LOW	NONE	NONE

Figure 3 presents the relationship between the *Test Environment* and the *Low Level Environmental Factors*. Note that "Low" and "High" *Overall Complexities* are achieved by single combinations of "Low" "Low" and "High" "High" *Feature Densities* and *Feature Complexities*, respectively. "Low/Medium" and "Medium/High" *Overall Complexity* can be obtained by two combinations of *Feature Density* and *Feature Complexity*. "Medium" *Overall Complexity* is achieved by three unique combinations. Since the *Lab* environment is heavily controlled by the *Evaluators* and it's usually desired to obtain specific *Technical Performance* data during the technology's early stages of development, it's unlikely that the *Overall Complexity* will exceed the "Medium" level. Note that it is possible to obtain *Utility Assessment* data in the *Lab*, but this *Test Environment* limits the type and range of qualitative data that can be captured since the *Lab* is not indicative of the *Actual* Environment. The *Simulated* and *Actual* environments are capable of producing the full range of *Overall Complexities* where the significant difference between the two is that the *Evaluators* have some measure of control over the parameters and variables present within the *Simulated* environment whereas the *Evaluators* have no control over test parameters and variables within the *Actual* environment.

It is also critical to note that the *Feature Density* and *Feature Complexity* ranges from "Low" to "Medium" to "High" correspond to global values across a specific *Test Environment*. For instance, the global *Feature Density* of an environment may be classified as "Medium" yet one local spot in the environment could have a large cluster of features indicating a "High" local *Feature Density*. Likewise, another spot within this same *Test Environment* could be sparsely populated so its local *Feature Density* could be classified as "Low." Altogether, the global *Feature Density* of the entire *Test Environment* is still "Medium."

Matching the appropriate *Explicit Environmental Factors* of the ARL CTA testing effort based upon previous test events is not trivial. Since NIST personnel have limited prior algorithm testing, it's difficult to state how the *Feature Density, Feature Complexity*, and *Overall Complexity* compare in the current test venue from prior testing environments. Also, since this technology will be integrated onto a greater system for its intended usage, it's difficult to state what the *Actual* environment will be, especially since the greater system(s) are somewhat unknown at this point. This highlights another challenge in accurately defining this blueprint category. Should *Explicit Environmental Factors* be referenced from the same test types with the technology at comparable states or do values of *Explicit Environmental Factors* range across all test types? In the case of the ARL CTA testing effort, the former would mean that "Low" *Feature Complexity* in the *Lab* is much lower and vary different to "Low" *Feature Complexity* in a *Simulated Environment*. The latter would mean that "Low" *Feature Complexity* in the *Lab* is comparable to "Low" *Feature Complexity* in a *Simulated Environment*.

The MOUT *Simulated Test Environment's Explicit Environmental Factors* that the ARL CTA/NIST team most recently evaluated the technology could be classified as having an *Overall Complexity* of "Medium" where *Feature Density* and *Feature Complexity* were both globally "Medium." This determination is based upon the overall consideration of the number of pedestrians within their environment, their motion paths, the number and type of fixed obstacles, number of lanes and other ambient features and/or obstacles. However, looking at specific artifacts and personnel activities within the environment, a case could be made that local *Feature Density* ranged from "Low" to "Medium" since multiple personnel were close proximity to one another in some spots while other personnel stood by themselves in other spots. Comparably, it can be stated that local *Feature Complexity* also ranges from "Low" to "Medium" considering that there were various environmental features present including several rectangular buildings and about a dozen *Participants*.

Further exploration needs to be completed on the exact method(s) to determine global and local complexities and densities both in general and specific to the CTA/NIST example. Additional time will be spent with the CTA/NIST test designers to obtain a greater understanding of their specific wants regarding the features and obstacles specifically placed in the environment.

The authors envision refining these blueprint specifications and solidifying the issue raised regarding if levels of *Explicit Environmental Factors* range within each *Test Environment* or across all *Test Environments*. Additionally, the authors will determine if the MRED framework will identify specific *Feature Density* and *Feature Complexity* levels, leading to unique *Overall Complexities* or if the framework will simply specify the *Overall Complexity* allowing the designer some freedom to specify *Feature Densities* and *Complexities* based upon the relationships identified in Figure 3.

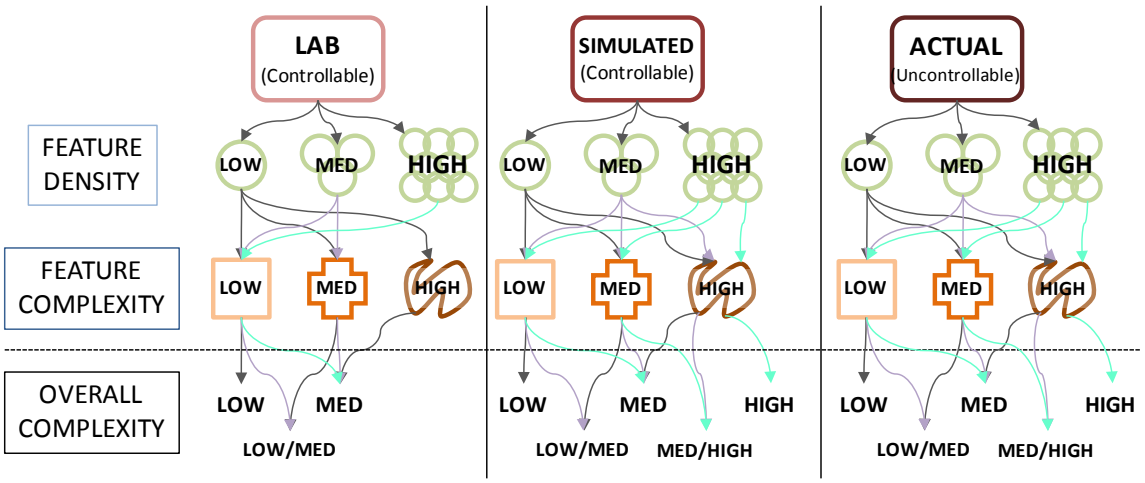

Figure 3. Relationship between the Test Environment and its Low Level Environmental Factors

3.4.3 Data Collection Methods

Technology-specific and/or customized *Data Collection Methods* are used to capture experimental and ground truth data depending upon the technology being evaluated and the environment it's being tested. *Data Collection Methods* are classified as being an observation device placed at a specific location in the environment to collect experimental test and/or ground truth data. No matter the type of tools are being used, *Data Collection Methods* include several factors that influence the techniques being employed. These two factors include *Mode* and *Collector Location* and are discussed further.

- *Mode* – This factor refers to the nature of the *Data Collection* Method that will be employed. Specifically, there are two different types of *Modes* that can collect data.
  1) *Automated Modes* involves collecting data with a calibrated technology that is independent of the system undergoing testing. An example of this would be using a radar gun to determine the speed of a vehicle.
  2) *Manual Modes* features an *Evaluator* actively managing a calibrate technology or collecting data by hand. Either way, *Manual Modes* are impacted by the *Evaluator* and are subject to human error. An example of *Manual Modes* would be a human starting and stopping a stop-watch to determine the time it takes a vehicle to go from point A to point B. Note that an evaluation can include *Data Collection Methods* of both *Automated* and *Manual Modes.*
- *Collector Location* – This factor refers to where the *Data Collection Methods* are located relative to the technology under test. The different perspectives include (i.e. physical locations from which observations are taken) include:
  1) From that of the technology (subject of the testing) – For those data collection tools used from this perspective during the evaluation, it's important that they are as discreet as possible so they do not interfere with the technology's functions.
  2) From that of a *Tech-User* – Depending upon the exact nature of the test and the type of data being collected (quantitative vs. qualitative metrics) it may be imperative for the *Data Collection Methods* to be as unobtrusive as possible so as not to influence the *Tech User* as they are using the technology.
  3) From that of a *Team Member* – Although the same types of data can be collected from this perspective as compared to that of the *Tech-User,* the captured data is distinct. The *Team Member(s)* have the ability to provide feedback about not only their perceptions of the technology's effectiveness, but also their perception of how it's impacting the *Tech-User* in their ability to complete their task, mission objective, etc.
  4) From that of a *Participant* – Data collected from this perspective is also distinct from that collected by the *Tech Users and Team Members*. The evaluation team gets insight from individuals who usually have the least familiarity with the technology. It should also be noted that *Participants* rarely support *Data Collection Methods* used during the evaluation since this can be perceived by the *Tech-Users* and *Team Members* as part of the evaluation scenario and not something that is purely for the evaluation.
  5) From the environment – Data collected from this perspective usually includes sensors automatically collecting data and/or evaluation team members manually collecting data from various points within the test environment.

Table 5 shows several example *Data Collection Methods* from both *Automated* and *Manual Modes* across the five *Collector Locations.*

The CTA/NIST team deployed numerous *Automated Data Collection Methods* from the *Collector Location* from within the test environment. Specifically, an Ultra-Wideband (UWB) tracking system is deployed to capture position ground truth data of the test vehicle, key environmental features and pedestrians (*Participants*) within the testing environment [4]. Data is used to capture quantitative *Technical Performance* data of the sensors and algorithms in order to generate the necessary evaluation metrics. A filter, devised by evaluation personnel, is incorporated into the raw UWB tracking data to minimize the impact of any potential data collection errors resulting from artifacts within the environment or from the UWB technology, itself. Additionally, numerous cameras are setup throughout the environment to automatically capture additional position data of test vehicle, key environmental features and *Participants.*

The experimental data is composed of each algorithm reporting detection information such as positions and velocities of the humans at the end of each CTA algorithm cycle. Some of the underlying assumptions for the outputs of the algorithms included:

- Only obstacles seen and classified as human were reported.
- Unique identification numbers were assigned to individual algorithm detections within a run.
- Algorithms demonstrated tracking of an individual by maintaining the same ID in successive frames.

Since the fixed test area was instrumented to capture ground-truth data, all detections were excluded if they occurred outside the test area. The correspondence algorithm found the correspondence between the detections and the ground-truth based on location and time stamp. Detections were compared with all the ground-truth objects on the course to attain the desired metrics [2].

Table 5. Example Data Collection Methods from various Collection Perspectives and Modes

	DATA COLLECTION METHODS: Examples	
COLLECTION PERSPECTIVE	*MEANS - AUTOMATED*	*MEANS - MANUAL*
From the Technology	Sensor and/or tool collecting technical information during the evaluation	Evaluator capturing data with sensors, tools
	Output of log files following the evaluation	Evaluator making notes of behavior
From the *Tech User*	Sensors (e.g. helmet camera) attached to the *Tech User* that collect data during testing	Surveys prior to and/or following the evaluation(s)
		Interviews prior to and/or following the evaluation
		Verbal and/or physical feedback provided by the *Tech User* during the evaluation (e.g. thumbs-up or thumbs-down at key way points)
From the *Team Member*	Sensors (e.g. microphone) attached to the *Team Member* that collect data during testing	Surveys prior to and/or following the evaluation(s)
		Interviews prior to and/or following the evaluation
		Verbal and/or physical feedback provided by the *Tech User* during the evaluation (e.g. thumbs-up or thumbs-down at key way points)
From the *Participant*	NONE (usually)	Surveys prior to and/or following the evaluation(s)
		Interviews prior to and/or following the evaluation
From the Environment	Sensors (e.g. radar gun, thermal camera, motion detector) setup throughout the environment that collect data during testing	Evaluation personnel stationed in various parts of the environment taking notes and/or manually using a sensor and/or tool to collect data

3.4.4 Personnel – Evaluators

There are three classes of evaluation personnel that are necessary to ensure that the evaluation proceeds accordingly to plan and that the necessary data is captured to evaluate a technology's performance. They fall into the three classes below:

- *Evaluators: Data Collectors* – These *Evaluators* are responsible for either setting up/implementing automated collection methods and/or performing manual collection methods. This class of *Evaluators* is also responsible for collecting experimental data directly from the technology at the conclusion of each test scenario (as necessary).
- *Evaluators: Test Executors* – These *Evaluators* are responsible for initiating the test including instructing *Participants* on when to engage in their specified activities within the environment.
- *Evaluators*: *Safety Officers* – These *Evaluators* are solely responsible for ensuring the safety of all personnel within the *Test Environment* along with protecting the technology and the environment, itself.

Depending upon the nature of the technology being evaluated and the range of *Data Collection Methods* employed, it's possible that some *Data Collectors* may also be *Test Executors* and vice versa. Although safety is everyone's responsibility on a test site, including *Data Collectors* and *Test Executors*, *Safety Officers* have no other role other than ensuring a safe test. The exact responsibilities and number of each of these personnel is heavily dependent upon the size and scope of the evaluation.

The ARL CTA/NIST testing featured both *Data Collectors* and *Test Executors* facilitating the test exercises. Specifically, the *Data Collectors* included personnel responsible for deploying and calibrating the UWB tracking system and cameras prior to the test event. These same personnel were also responsible for managing the UWB tracking system and cameras during test to ensure it was operating within normal limits. Numerous *Test Executors* also played a significant role in the evaluation. This personnel class included one individual who signaled the start and conclusion and another individual that signaled the *Participants* to walk in their prescribed paths. Additionally, another *Test Executor* was employed to signal when the vehicle should begin its motion and when the sensors and algorithms under test should be engaged. This test exercise featured numerous *Safety Officers* stationed throughout the environment. Additionally, several *Safety Officers* were positioned at key locations along the test environment's perimeter to prevent non-evaluation personnel or vehicles from entering.

4. CONCLUSION

The MRED model's new blueprint elements have shown they can be applied to a current technology through test plan matching with the ARL CTA test plans. This has already highlighted some areas to address in this continuing work. The next steps for the MRED model are to identify the remaining evaluation blueprint pieces and continue to validate its design against a technology whose own tests were inspired by other successful methods, such as the SCORE framework. Once the entire blueprint has been specified, the three input categories will be addressed in detail. Since each input is nondeterministic in nature (based upon human preference, an unknown technology state, or uncertain resource availability), uncertainty will be factored. The inner workings of the framework

will then be outlined and devised, first assuming certain inputs, and then uncertain data.

With MRED leveraging some of the success of a previous evaluation framework, MRED presents expanded capabilities in the detailed evaluation blueprints it prescribes along with the defined relationships among them. It is envisioned that MRED will be an invaluable tool in devising comprehensive technology test plans of emerging and advanced intelligent systems allowing evaluation designers to be more effective and efficient in producing and implementing the appropriate tests.

5. ACKNOWLEDGMENTS

The authors would like to thank Harry Scott, the NIST project leader for the ARL CTA testing effort, for his continued support throughout this work. Further, the lead author would like to thank Craig Schlenoff of NIST for his support and encouragement in this effort.

6. 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] 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.*

[3] 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).

[4] Fontana, R., Richley, E. and Barney, J., 2003, "Commercialization of an Ultra Wideband Precision Asset Location System," In *2003 IEEE Conference on Ultra Wideband Systems and Technologies*.

[5] Galvin, C., 2010, "Gulf Oil Spill: Responding with Robots," *Robotics Trends*. DOI = http://www.roboticstrends.com/service_robotics/article/gulf_oil_spill_responding_with_robots

[6] 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..

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

[8] 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.

[9] 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*.

[10] 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.*

[11] 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*.

[12] 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.*

[13] 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.

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

[15] 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.

[16] 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.

[17] 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*.

[18] Weiss, B.A., Schmidt, L.C., Scott, H.A., and Schlenoff, C.I., 2010, "The Multi-Relationship Evaluation Design Framework: Designing Testing Plans to Comprehensively Assess Advanced and Intelligent Technologies," *Proc. of the ASME 2010 International Design Engineering Technical Conferences (IDETC) – 22nd International Conference on Design Theory and Methodology (DTM).*

[19] 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*.

[20] 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.*