# A Taxonomy for the Evaluation of Training Simulations and Environments

**Martin S. Goodwin, Lauren Reinerman-Jones, Scott Harris**
**University of Central Florida, Institute for Simulation and Training (UCF IST)**
**Orlando, Florida**
**mgoodwin@ist.ucf.edu, lreinerm@ist.ucf.edu, sharris@ist.ucf.edu**

**Alexander Arrieta**
**United States Marine Corps Training and Education Command**
**Quantico, Virginia**
**Alexander.arrieta@usmc.mil**

## ABSTRACT

As the complexity of training events continues to evolve, training program effectiveness and the capabilities of simulation systems to support and optimize training outcomes becomes an increasingly critical concern. Historically, these concerns have been addressed through the use of traditional training evaluations. However, traditional evaluation methodologies do not adequately capture the complete range of efficacy factors that exist in modern training simulations. This paper addresses this gap by outlining a training evaluation taxonomy that identifies two main training evaluation components: the human element and the systems element. The human element includes assessment of the training tasks, objectives, and overall instructional design that drives the training experience. The human element of training evaluation is often referred to as a training effectiveness evaluation (TEE) and frequently includes measures of trainee perceptions, behaviors, and performance. The systems element of training evaluation involves an assessment of the instructional interfaces, technologies, and environments used to support and facilitate the performance of training tasks and requirements. It includes a review of the technology configuration used to support training, an attribute analysis of the training system, and documentation of operability/interoperability issues. This systems evaluation, known as a technology capability assessment (TCA), identifies system capabilities and limitations for training specific learning objectives when used in either stand-alone or distributed training configurations. This taxonomy helps guide training evaluation efforts by focusing and aligning assessment activities with desired assessment outcomes to provide key information to stakeholders and decision makers on the efficacy of mission critical training systems.

## ABOUT THE AUTHORS

**Martin S. Goodwin, Ph.D.** is a Research Associate at the University of Central Florida Institute for Simulation and Training. His research focuses on dynamic instructional systems, simulation and gaming technology integration, and evaluation methodologies to improve learning, engagement, and retention in virtual environments.

**Lauren Reinerman-Jones, Ph.D.** is the Director of Prodigy, which is one lab at the University of Central Florida's Institute for Simulation and Training, focusing on assessment for explaining, predicting, and improving human performance and systems.

**Scott Harris** is a Faculty Research Associate at the University of Central Florida Institute for Simulation and Training focusing on Department of Defense integration and research. A retired Marine Corps aviator, Scott leads several high-profile training assessment projects for the United States Marine Corps.

**Alexander Arrieta** is the Verification, Validation and Accreditation (VV&A) Lead for Marine Corps Synthetic Training Environment (MCSTE) Branch, Training and Education Capabilities Division, Training and Education Command. Alex is responsible for organizing, coordinating, and executing a comprehensive VV&A program for Marine Corps training Modeling & Simulations systems.

# A Taxonomy for the Evaluation of Training Simulations and Environments

**Martin S. Goodwin, Lauren Reinerman-Jones, Scott Harris**
**University of Central Florida, Institute for Simulation and Training (UCF IST)**
**Orlando, Florida**
**mgoodwin@ist.ucf.edu, lreinerm@ist.ucf.edu, sharris@ist.ucf.edu**

**Alexander Arrieta**
**United States Marine Corps Training and Education Command**
**Quantico, Virginia**
**Alexander.arrieta@usmc.mil**

## INTRODUCTION

Training technologies have become more complex, requiring a greater emphasis on the proper alignment of training tasks with training capabilities and resources to support sustained performance of training requirements. The key to achieving the greatest levels of training optimization is through timely evaluation of mission critical training systems. In traditional contexts, training evaluations typically follow a unidimensional model that focuses on a hierarchy of levels to assess the value of training in terms of a defined set of objectives (Kirkpatrick, 1994). This type of evaluation has utility where the scope of the evaluation is limited to determining the effectiveness of training in accomplishing its desired objectives. However, training simulations and simulated environments introduce new variables, and thus new questions, related to training effectiveness. Evaluating the effectiveness of training in today's simulation domains typically extends beyond assessing the value of training. The imperative to determine if training is meeting its desired objectives still exists, however, this is now generally part of a much larger evaluation goal that encompasses the need to inform decisions concerning how, what, when, and where training simulations will be used to meet specific training requirements (Norman, Dore, & Grierson, 2012; Nyssen, Larbuisson, Janssens, Pendeville, & Mayné, 2002; Rothstein & Selman, 2015; Scerbo & Dawson, 2007). Optimizing training value is also an issue. Aligning training objectives and tasks with the most cost-efficient training environment while maintaining required levels of training effectiveness is a constant concern for stakeholders. As such, evaluation questions now become much broader in scope. Instead of simply determining if training is meeting its desired goals, decisions need to be informed concerning how the array of available training options will be leveraged to optimize a solution to an existing training requirement. These decisions are typically based on factors unique to training simulations, such as levels and types of fidelity, the affordances of instructional interfaces, and the dynamics of the environments themselves.

## A NEW TRAINING EVALUATION TAXONOMY

The dynamic nature and technological range of training simulations and virtual environments requires a new, more holistic training evaluation paradigm to comprehensively assess training capability and effectiveness. To address this gap, we propose a new training evaluation framework called the taxonomy for holistic evaluation and training assessment, or THETA. This taxonomy captures the two primary elements critical to comprehensive evaluation of training simulations and virtual environments, the human element and the systems element.

The human element of training evaluation is often referred to as a training effectiveness evaluation (TEE) and frequently involves the use of technologies for enhanced or augmented training, but does not always use simulators or simulation technologies. TEEs include measures of trainee perceptions, behaviors, performance, and more recently physiological responses during and after training in a simulated environment. These measures are valuable for a single training scenario with simulation technologies, but are stronger when repeated over the course of time for the same person on that given training because rate of learning and retention can be tracked more accurately. The gold standard is to assess the person in the training environment and for that same person to be assessed in operation. The systems element of training evaluation involves an assessment of the instructional artifacts, interfaces, and technologies used to support and facilitate the performance of training tasks. It includes a review of the technology configuration used

to support training, an attribute analysis of the training system based on the systematic team assessment of readiness training (START) methodology, and documentation of operability/interoperability issues. This systems evaluation, known as a technology capability assessment (TCA), informs decision makers about the evaluated system(s) capabilities and limitations for training specific learning objectives when used in either stand-alone or federated mission configurations. The THETA training evaluation taxonomy is illustrated in Figure 1.

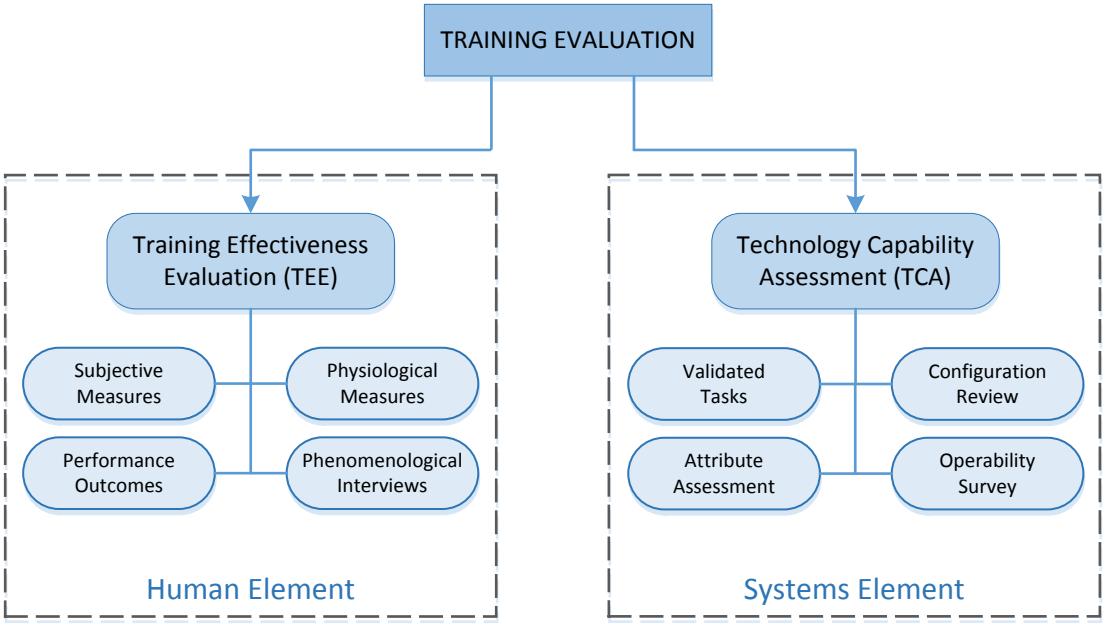


**Figure 1: Taxonomy for Holistic Evaluation and Training Assessment (THETA)**

**Training Effectiveness Evaluations: Assessing the Human Element**

A training effectiveness evaluation is the process of determining whether a training program is meeting its intended goals (Coultas, Grossman, & Salas, 2012). TEEs are conducted to gather data relating to specific aspects of a training program. This data is then used to verify if the training is meeting its desired intent and to assess the overall value of instruction. As TEEs have traditionally been developed from the field of education and more specifically, instructional design, TEEs have included both formative and summative evaluations (Rothwell & Kazanas, 2008). Formative evaluation activities are typically conducted throughout the training development process to verify and validate instructional sequence and content. Formative evaluations typically inform improvements in the design and delivery of instruction (Rothwell & Kazanas, 2008). Summative evaluations are typically performed after the training program is implemented. Burkett (2002) notes that summative evaluations are usually conducted from a performance improvement perspective and are used to:

- determine if a training program accomplished its objectives;
- determine if a performance gap was closed or narrowed as a result of the training;
- determine if the training met the intended goals;
- determine the benefit/cost ratio of a training program; and provide data to justify the implementation, expansion, reduction, or elimination of training programs and activities.

**Traditional TEE Approaches**

Whether formative or summative, the most popular and widely used methods for performing training evaluations are based on Kirkpatrick's four-level training evaluation model (Anguinis & Kraiger, 2009; Bates, 2004; Mathieu & Martineau, 1997). This model evaluates training effectiveness through an assessment of four individual levels (Kirkpatrick, 1959, 1976, 1994; see Figure 2).


**Figure 2: Kirkpatrick's Four-Level Training Evaluation Model**

- Level 1: Reaction – Evaluates trainees' reactions to the training event.
- Level 2: Learning – Evaluates changes in trainees' knowledge, skills, attitudes, and abilities as a result of the training event.
- Level 3: Behavior – Evaluates the change in behavior in trainees from the training context to the performance context to determine training transfer and application.
- Level 4: Results – Evaluates the degree to which specific targeted outcomes have been achieved.

The original purpose of the Kirkpatrick model was to gain information on the value of training programs to help determine instructional improvements, decide if a program should be continued, and justify the existence of training departments as contributors to the goals of an organization (Kirkpatrick, 1994). As such, it follows a traditional evaluation methodology and is most effective in assessing static training contexts (i.e., training contexts lacking the dynamic nature and capabilities of technology-rich, immersive learning environments) to determine if instructional improvements are required. Evaluations conducted in simulation training contexts are typically less concerned about *improving* a single training program and more concerned about *proving* the efficacy of specific individual factors, such as different types of fidelity, that influence training effectiveness. This focus on proving instead of improving necessitates the use of a TEE approach based on a research methodology instead of a standard evaluation methodology.

**Enhanced TEE Approach**

A more comprehensive TEE approach that addresses the limitations of the Kirkpatrick model, called assessing simulated systems empirically for training, or ASSET, leverages an evaluation paradigm better aligned with the purpose and objectives of TEEs in modern, technology-enabled training environments (Goodwin, Reinerman-Jones, Goldiez, & Crapanzano, 2017). ASSET draws on the tools and techniques of human performance assessment, instructional science, and neurophenomenology to establish a multidimensional, interdisciplinary perspective to performing TEEs. This approach increases the breadth of evaluation efforts to more fully capture the range of factors that contribute to training effectiveness in dynamic, interactive simulation training environments. ASSET follows the procedures and rigor of a research methodology, with some slight modification to optimize its use to conduct TEEs in simulation training environments.

The ASSET approach begins with an identification of the scope and objectives of the evaluation. This is an essential part of the process, as it frames the questions that will establish the parameters for performing the TEE. This is the distinguishing characteristic of the ASSET approach. Instead of limiting the evaluation to questions relating to trainee reactions, changes in learning or behavior, or specific organizational results, the ASSET approach facilitates TEEs that address a broad range of questions. This is achieved by leveraging the dynamic nature of simulation environments to develop experimental scenarios that target specific questions of interest. These scenarios may incorporate a suite of performance metrics, such as task completion, timed events, or cognitive decision points, to develop custom performance rubrics and further define evaluation questions.

The ASSET approach also outlines an interdisciplinary set of empirically validated measures that contribute to training effectiveness. These measures are aligned within the disciplinary areas of psychology, physiology, and phenomenology. Individuals may show distinct yet unique responses, both psychologically and physiologically, when

appraising, developing strategies for, or performing a task. Single metrics or measures may provide misleading assessments (Matthews, Reinerman-Jones, Barber, & Abich, 2015; Saxby, Matthews, Warm, Hitchcock, & Neubauer, 2013). Therefore, the ASSET approach employs multiple measures to help identify unique patterns of responses during the TEE process. The remaining steps of the ASSET approach follow a traditional research methodology.

The psychological, performance, physiological, and phenomenological data captured using the ASSET approach facilitates a broader and deeper analysis of training effectiveness measures (Reinerman-Jones, Goodwin, Wismer, Goldiez, & Crapanzano, 2017). The ASSET approach is illustrated in Figure 3.

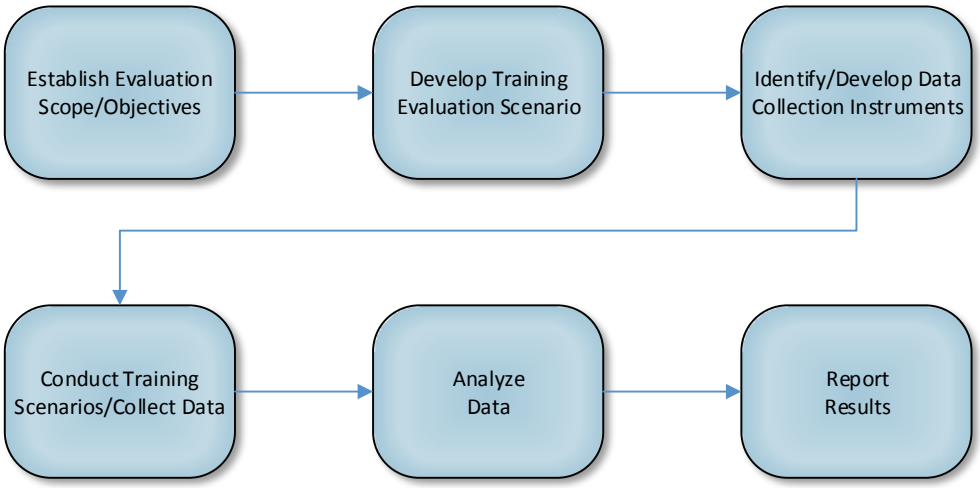


**Figure 3: ASSET Evaluation Approach**

**Technology Capability Assessment: Assessing the Systems Element**

The systems element of training includes the technologies and user interfaces that support and facilitate the performance of training tasks. It consists of the integrated hardware, software, and instructional artifacts that form a training system. Assessing the efficacy of training system technologies requires an approach that is different from the training effectiveness evaluation described in the previous section. Instead of emphasizing the trainee-centric measures of TEE approaches, training system assessments must focus on the technical capabilities of the system to support training requirements. The ability of the training system to provide sensory inputs (stimuli) to the trainee to establish an operational context and influence task performance relative to training objectives should be the primary focus of the assessment effort. The sensory inputs, or stimuli, provided by the training system are known as its attributes. Assessment measures should concentrate on the degree to which the attributes of the training system support the performance of individual training tasks, and thus meet training requirements. This type of training system technology assessment is not uncommon, however, it is typically conducted using TEE approaches. Following the trainee-centric measures used in these approaches, training system technical capabilities are usually assessed in terms of trainee perceptions, behaviors, and performance instead of through measures which more accurately align training system technological capabilities and attributes with training tasks and requirements.

This misapplication of TEEs to evaluate training system technological capabilities may be due to a lack of other viable assessment approaches. The training evaluation literature abounds with TEE concepts, approaches, and models that provide trainee-centric assessment methods (Bushnell, 1990; Kaufman, Keller, & Watkins, 1995; Kirkpatrick, 1959, 1976, 1994; Worthen & Sanders, 1987). However, very little information exists that focuses on the assessment of training system technologies. The literature that is available in this area either maintains a heavy reliance on trainee perceptions and performance (e.g., Livingston, Dyer, & Swinson, 2005) or effectively removes training tasks from the assessment paradigm (e.g., Fu, Jensen, & Hinkelman, 2008). Further, an exhaustive literature review found no evaluation approaches that specifically assess the alignment of training system technical capabilities and attributes

with training tasks and requirements, exposing a sizable gap in training evaluation practice. The technology capability assessment (TCA) addresses this gap by establishing a systematic, data-driven methodology to evaluate the systems and technologies that support training events. Developed by the Prodigy Lab at the Institute for Simulation and Training at the University of Central Florida, the TCA methodology provides a comprehensive approach for documenting, assessing, and reporting training system capabilities. TCAs are performed in both stand-alone and distributed mission training environments, providing valuable assessments of training assets in both single system and system-of-systems configurations. TCAs consist of five components: task list development, review of the training system configuration, analysis of training system attributes, an operability/interoperability survey, and a training system TCA report (see Figure 4).

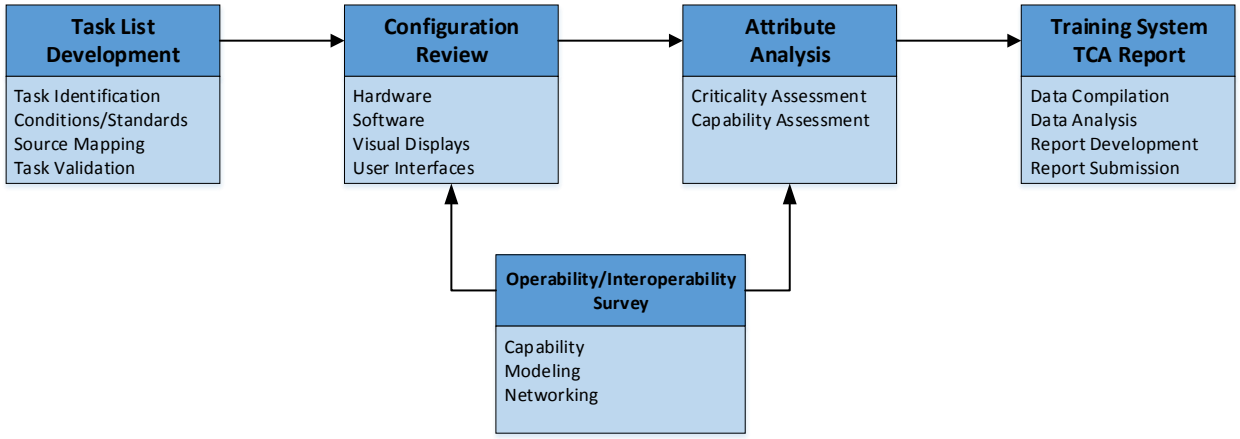


**Figure 4: Technology Capability Assessment (TCA) Process**

**Task List Development**
Tasks are the primary drivers of assessment efforts in the TCA process. A task is defined as any activity conducted by an individual or a team in the actual performance environment that is necessary to accomplish a specific job or mission. A task may be knowledge-based or skill-based and may support individual or collective actions. The first step in preparing for a TCA is the development of a comprehensive list containing a set of mission essential and key tasks associated with specific operations or tactics. These tasks ultimately form the basis for the overall TCA.

Task list development involves an identification of the training requirements and objectives associated with the training system that is the subject of the assessment. Specific tasks are derived from publications; manuals; instructions; tactics, techniques and procedures (TTPs); and other relevant sources. The task list is validated and finalized in consultation with subject matter experts (SMEs) from relevant operational and training communities.

**Configuration Review**
The configuration review captures information unique to the training system that is the subject of the assessment. This review documents specific information on the training system's hardware, software, visual display system, audio system, and user interfaces. This information helps establish a system technical baseline while also capturing configuration changes that may have been performed to address operability/interoperability issues.

**Attribute Analysis**
Attribute analysis is the heart of the TCA methodology. It builds upon the proven systematic assessment of team readiness training (START) process developed by the Naval Air Warfare Center, Training Systems Division (NAWC-TSD) for aviation simulation training. This process has since been adopted and further refined for the assessment of ground simulation training by the United States Marine Corps (USMC), Training and Education Capabilities Division (TECD) and Program Manager, Training Systems (PMTRASYS). As a result, attribute analysis has become a robust, generalizable tool for assessing the capability of training systems to enable and support training tasks in stand-alone or distributed configurations (Dunne, Harris, Arrieta, Tanner, Vonsik, Lalor, & Muir 2017).

Attributes are defined as those elements and stimuli provided by the environment that support task performance. During attribute analysis, tasks are aligned to training system attributes and the degree to which those attributes enable and support task performance is assessed. A listing of sample attributes is contained in Table 1.

**Table 1. Sample Attributes**

| Attribute Category | Attribute | Attribute Definition |
|---|---|---|
| Physical Look and Feel | Appearance (physical properties) | Ability to convey an accurate and realistic representation of an object and its physical properties and/or distinguishing characteristics - such as shape, size, color, mass, or relative position with respect to other objects. |
| | Tactile Feel (touch sensation) | Ability to convey an accurate and realistic representation of the sensations produced by physical contact (i.e., physical feel) of an object, instrument, or natural element usually located in the immediate environment or vicinity. |
| Functionality | Haptic Cues (kinesthetic response) | Ability to convey an accurate and realistic "response" sensation (feedback) when touching or interacting with the ground systems, instruments and controls, etc. This can include the relative weight (portability) of the system, object or piece of equipment (or pressure exerted) when pulling, pushing or lifting it. |
| | Systems Response Interaction | Ability of the system to provide a realistic and appropriate interoperability - i.e., system reaction in response to user input, or where input or activity with one system (controls,, instruments, etc.) influences the reaction, output or display of another system, control or instrument. |
| Auditory | Environmental & Battle Sound | Ability to convey discernable and indiscernible sounds occurring in the environment - whether natural or man-made - including variable battle munitions sound. |
| | Sound Bearing | Ability to convey variable sound tone, volume, and source location in order to be able to determine the direction of the sound. |
| | Audible Systems Signals (devices) | Ability to convey realistic non-verbal audio sounds and tones where the tenor, frequency, tone, regularity, pitch and/or volume provide meaningful and specifically interpretable information, signals or alerts. |
| Visual | Static Visuals (projected) | Ability to convey accurate and realistic representations of stationary environmental objects, terrain, topography, and scenery. These objects and scenery are generally reproduced via screen projections and NOT actual objects within one's reach. |
| | Active Visuals (projected) | Ability to convey accurate and realistic representations of moving objects in the environment (ground and sky) with an appropriate motion fidelity, motion speed, acceleration, and motion trajectory. |
| | Aero Models | Ability to convey accurate and realistic aero models for airborne aircraft and objects (munitions, missiles, rockets, airborne threats, etc.) including airspeeds, trajectories, turns, dives, climbs, maneuvers, etc. for single and multiple objects, munitions or aircraft. |
| | Depth Perception | Ability to convey accurate and realistic representations of realistic distances or changing distances of objects in the environment (on the ground or in the air). |

Attribute analysis is a data collection effort accomplished in two phases that assesses task to attribute criticality and capability. Data input is provided by SMEs from the relevant operational and training communities. A criticality assessment is performed first. During this effort, training analysts guide SMEs in evaluating, task by task, how critical the presence of specific attributes are to execute specific tasks. Each task/attribute combination is rated based on a five-point scale defined and described in Table 2. It's important to note that this criticality assessment is not specific to any particular simulation training system or exercise. It therefore reflects the criticality of the associated attribute in the live performance context.

**Table 2. Criticality Ratings and Definitions**

| Rating | Attribute Criticality | Attribute Criticality to Task Performance |
|---|---|---|
| **5** | Absolutely Critical | Task cannot be executed without this attribute. |
| **4** | Critical | Attribute is critical, contributing to important cues to task execution. |
| **3** | Important | Attribute is important and contributes to task execution, but work-around is acceptable. |
| **2** | Nice but not important | Attribute is nice to have but peripheral and not essential to task execution. |
| **1** | Irrelevant | Attribute is irrelevant or not applicable and contributes nothing to task execution. |

The second phase of the attribute analysis is a capability assessment. This phase is similar to the criticality assessment, except that it focuses on the capabilities of a specific training system. Guided by training analysts, SMEs evaluate the capability of each training system attribute, task by task, to enable and support performance of specific tasks. Each task/attribute combination is rated based on a five-point scale defined and described in Table 3.

**Table 3. Capability Ratings and Definitions**

| Rating | Attribute Capability | Device Capability to Enable Task Performance |
|---|---|---|
| **5** | Fully Capable | Device is fully capable of providing attribute to support task performance with little or no capability gaps and no departure from realism. No compensation needed to support task execution. |
| **4** | Effectively Capable | Device effectively provides attribute to support task execution with minor/annoying capability gaps and some departure from realism. Minimal compensation needed to support task execution. |
| **3** | Borderline Capable | Device is borderline capable of providing attribute to support task execution with moderate capability gaps and significant departure from realism. Considerable compensation needed to support task execution. |
| **2** | Marginally Incapable | Device is marginally incapable of providing attribute to support task execution with significant capability gaps and very little realism. This severely diminishes the device's capability of supporting task execution. |
| **1** | Completely Incapable | Device is completely incapable of providing attribute to support task execution. |

Attribute analysis data collection is followed by computational determination of criticality and capability scores based on the provided data.

The attribute analysis provides the following information:

- Specifies training device attributes (sensory input provided by the training device to the user to provide operational context and influence task performance) that are required to effectively support performance of tasks associated to specific training events.
- Determines which training device attributes provide sufficient simulation fidelity for the training environment.
- Identifies deficiencies in training device attributes which require improvement to support training tasks and requirements (e.g., visual or auditory stimuli may need to be improved to better support training task performance).

**Operability/Interoperability Survey**
The operability/interoperability survey is a qualitative assessment that documents system functionality and performance in the areas of modeling, networking, correlation, and capability. An important component of the TCA methodology, it is used during configuration review and attribute analysis to document any workarounds required to ensure proper system or system-of-systems functionality in support of training requirements. This captures and preserves systems information to be included in operations and planning documents to inform future system upgrades and engineering changes.

**Training System TCA Report**
TCA reports for individual training systems document and disseminate system configuration data, attribute analysis results, and operability/interoperability survey information. This report provides decision makers and training stakeholders with specific, actionable information on training system capabilities and limitations for specific training requirements and objectives. Through its detailed analysis to the task and attribute level, the TCA methodology identifies and reports specific areas for training system upgrades to improve training efficiency, advance training objectives, and ultimately save lives.

## SUMMARY

Evaluating the capabilities and training effectiveness of today's advanced simulations is often performed using methodologies that were not designed to assess dynamic, interactive environments. As a result, there is a pressing need for more relevant and comprehensive training simulation evaluation methods. This paper proposes a new training evaluation framework called the taxonomy for holistic evaluation and training assessment, or THETA. This taxonomy captures the two primary elements critical to comprehensive evaluation of training simulations and virtual environments, the human element and the systems element. The human element includes assessment of the training tasks, objectives, and overall instructional design that drives the training experience. The systems element of training evaluation involves an assessment of the instructional interfaces, technologies, and environments used to support and facilitate the performance of training tasks and requirements. Overall, this taxonomy helps guide training evaluation efforts by focusing and aligning assessment activities with desired assessment outcomes to provide key information to stakeholders and decision makers on the efficacy of mission critical training systems.

## ACKNOWLEDGEMENTS

The authors wish to acknowledge the assistance and cooperation provided by MajGen Mel Spiese (ret) and the USMC MCTOG community, Matt Denny, Maj Jesse Attig, and Col Timothy Barrick.

## REFERENCES

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology, 60*, 451-474.

Bates, R. (2004). A critical analysis of evaluation practice: the Kirkpatric model and the principle of beneficience. *Evaluation and Program Planning 27*(2004), 341-347.

Burkett, H. (2002). *Evaluation: Was your HPI project worth the effort?* In G.M. Piskurich (Ed.), HPI essentials. American Society for Training and Development: Alexandria, VA.

Bushnell, D. S. (March, 1990). Input, process, output: A model for evaluating training. *Training and Development Journal, 44*(3), 41-43.

Coultas, C. W., Grossman, R. & Salas, E. (2012) Design, Delivery, Evaluation, and Transfer of Training Systems, in G. Salvendy (Ed.) *Handbook of Human Factors and Ergonomics, Fourth Edition*, Hoboken, NJ: John Wiley & Sons, Inc.

Dunne, R., Harris, S., Arrieta, A., Tanner, S., Vonsik, B., Lalor, J., & Muir, S. (2017). *Live, Virtual, Constructive Distributed Missions: Results and Lessons Learned.* Proceedings of the Interservice/Industry Training, Simulation and Education Conference 2017. Orlando, Florida.

Fu, D., Jensen, R., & Hinkelman, E. (2008). Evaluating Game Technologies for Training. In *Aerospace Conference, 2008.* Institute of Electrical and Electronics Engineers.

Goodwin, M.S., Reinerman-Jones, L., Goldiez, B.F., & Crapanzano, R.A. (2017). *An interdisciplinary approach to evaluating U.S. Army aviation training.* International Symposium on Aviation Psychology, 2017.

Kaufman, R., Keller, J.M., & Watkins, R. (1995). What works and what doesn't: Evaluation beyond Kirkpatrick. *Performance and Instruction, 35*(2), 8-12.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of American Society for Training and Development*, 11, 1-13.

Kirkpatrick, D. L. (1976). Evaluation of training. In R.L. Craig (Ed.), *Training and development handbook: A guide to human resource development*. New York: McGraw Hill.

Kirkpatrick, D. L. (1994), *Evaluating Training Programs: the Four Levels*. San Francisco, CA: Berrett-Koehler.

Livingston, S. C., Dyer, J. L., & Swinson, D. (2005). *A Training Technology Evaluation Tool*. Columbus, Georgia: Northrup Grumman Mission Systems.

Mathieu, J.E. & Martineau, J.W. (1997). Individual and situational influences on training motivation. In J. K. Ford, S.W. J. Kozlowski, K. Kraiger, E. Salas, & M. Teachout (Eds.), *Improving Training Effectiveness in Work Organizations*. New York: Taylor and Francis Group.

Matthews, G., Reinerman-Jones, L.E., Barber, D.J., & Abich, J., IV. (2015). The psychometrics of mental workload: Multiple measures are sensitive, but divergent. *The Journal of the Human Factors and Ergonomics Society*.

Norman, G., Dore, K., & Grierson, L. (2012). The minimal relationship between simulation fidelity and transfer of learning. *Medical education*, *46*(7), 636-647.

Nyssen, A. S., Larbuisson, R., Janssens, M., Pendeville, P., & Mayné, A. (2002). A comparison of the training value of two types of anesthesia simulators: computer screen-based and mannequin-based simulators. *Anesthesia & Analgesia*, *94*(6), 1560-1565.

Reinerman-Jones, L., Goodwin, M. S., Wismer, A.J., Goldiez, B.F., & Crapanzano, R.A. (2017). *Toward augmenting Army aviation collective training with game-based environments*. Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC 2017). Orlando, Florida.

Rothstein, B. D., & Selman, W. R. (2015). Evaluating simulation as a teaching tool in neurosurgery. *Virtual Mentor*, *17*(1), 33.

Rothwell,W. J., & Kazanas, H.C. (2008). *Mastering the Instructional Design Process: A Systematic Approach*, 4th edition. New York: JohnWiley& Sons.

Saxby, D. J., Matthews, G., Warm, J. S., Hitchcock, E. M., & Neubauer, C. (2013). Active and passive fatigue in simulated driving: Discriminating styles of workload regulation and their safety impacts. *Journal of Experimental Psychology: Applied, 19*(4), 287-300.

Scerbo, M. W., & Dawson, S. (2007). High fidelity, high performance? *Simulation in Healthcare, 2*(4), 224-230.

Worthen, B. R., & Sanders, J. R. (1987). *Educational evaluation*. New York: Longman.