AN INTERDISCIPLINARY APPROACH TO EVALUATING U.S. ARMY AVIATION TRAINING

Martin S. Goodwin
University of Central Florida, Institute for Simulation & Training
Orlando, Florida
Lauren Reinerman-Jones
University of Central Florida, Institute for Simulation & Training
Orlando, Florida
Brian F. Goldiez
University of Central Florida, Institute for Simulation & Training
Orlando, Florida
Robert A. Crapanzano
U.S. Army Program Executive Office for Simulation, Training, and Instrumentation (PEO STRI)
Orlando, Florida

The U.S. Army is seeking to update and expand its use of simulation-based aviation training to address operational and fiscal concerns that are driving the need for more efficient training solutions. This has created a need to evaluate whether lower-cost, game-based simulations may potentially augment higher-cost, traditional simulation-based training for specific aviation training tasks. However, current approaches to Training Effectiveness Evaluation (TEE) do not address the complete range of factors to adequately evaluate today's increasingly sophisticated simulation training environments. Leveraging recent research and drawing from the tools and techniques of human performance assessment, instructional science, and phenomenology, an interdisciplinary approach to performing TEEs is introduced and described in the context of evaluating UH-60A/L aviation collective mission training. This novel TEE approach optimizes a research-based evaluation methodology to more fully capture the range of factors that contribute to training effectiveness in interactive simulation training environments.

The United States continues to face uncertain and unprecedented threats around the world. Increasing acts of terror by both state and non-state actors, rising global instability, and the need to maintain readiness for both conventional and unconventional warfare are key strategic concerns. At the same time, technology innovations such as the expanding role of unmanned aircraft systems (UAS) and the emergence of the cyber-battlefield are changing the characteristics of modern warfare. Today's warfighters must be prepared to meet the challenges of highly dynamic, increasingly technological military operations. To help prepare warfighters to meet those challenges, the U.S. Army is seeking to update and expand its use of simulation-based aviation training. While the Army continues to rely on traditional simulation as a proven aviation training method, game-based simulation has become more sophisticated and may provide viable training options in some applications. The use of game-based simulation to augment traditional simulation-based training can potentially reduce costs, enhance return on investment, advance training objectives, and inform future training environment designs.

Operational imperatives are mandating training strategies that produce optimum levels of readiness for a wide range of mission scenarios. Simultaneously, fiscal concerns are driving the need for more efficient training methods. This need for optimized training can be addressed for the U.S. Army by investigating whether lower-cost, game-based simulations may potentially augment higher-cost, traditional simulation-based training for specific aviation training tasks. Such investigations are typically performed by conducting Training Effectiveness Evaluations (TEEs). The most popular and widely used methods for performing training evaluations are based on Kirkpatrick's Four-Level Training Evaluation Model (1959, 1976, 1994). However, the Kirkpatrick model does not adequately address the complete range of factors that exist in dynamic training simulations. Additionally, the model inherently limits the types of questions that need to be answered to effectively evaluate today's increasingly sophisticated simulation training environments. It also provides little guidance on how different simulated environments may be combined to meet evolving training requirements. This paper describes the structure of the Kirkpatrick model, the reasons for its popularity in the training community, and the contrast between its intended purpose and its use to address modern simulation training evaluation objectives. A novel, interdisciplinary approach to evaluating training effectiveness, called Assessing Simulated Systems Empirically for Training, or ASSET, is then introduced. ASSET addresses the limitations of TEE methods based on the Kirkpatrick Model by building on a methodology better aligned with the purpose of modern TEEs. The ASSET approach is then described in the context of a use case to evaluate whether game-based systems can potentially augment traditional simulation-based U.S. Army UH-60A/L Blackhawk helicopter collective training.

Training Effectiveness Evaluation Considerations

Kirkpatrick's Four-Level Training Evaluation Model (1959, 1976, 1994) seeks to evaluate training effectiveness through an assessment of four hierarchical levels (Figure 1).

- 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.

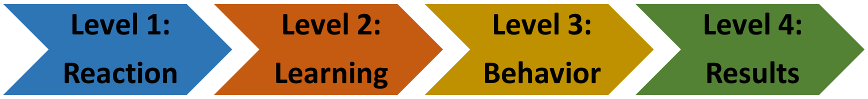

Figure 1. Kirkpatrick's Four-Level Training Evaluation Model

The popularity of the Kirkpatrick Model can be traced to a number of factors: 1) it provides a multi-level approach to training evaluation; 2) it organizes the complexities of training evaluation into four distinct areas; and 3) it simplifies outcome measures by reducing the number of variables involved in the evaluation analysis (Bates, 2004). The Kirkpatrick Model is used to conduct TEEs in many different training contexts, but its use to evaluate modern simulation training is problematic. The original purpose of the Kirkpatrick Model was to gain information on the *value* of training programs to help determine instructional improvements and decide if a program should be continued (Kirkpatrick, 1959). As such, it follows a traditional evaluation methodology and has utility in evaluation contexts where the intent is to determine whether the training is meeting desired objectives. In other words, the scope of the evaluation is limited to assessing a single training program in terms of the need it was designed to meet. Evaluating the effectiveness of training in today's simulation domains typically extends beyond this concern. While the imperative to determine if training is meeting its desired objective still exists, this is now generally part of a much larger evaluation goal that encompasses the need to inform decisions concerning how, what, when, and where simulation training will be used to meet specific training requirements. These decisions are typically based on factors unique to simulated environments, such as levels and types of fidelity, the affordances of instructional interfaces, and the dynamics of the environments themselves.

For simulation training then, TEEs are less concerned about *improving* a single training program and more concerned about *proving* the efficacy of specific individual factors 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. It is from this perspective that the interdisciplinary TEE approach called Assessing Simulated Systems Empirically for Training (ASSET) was deveoped.

Assessing Simulated Systems Empirically for Training (ASSET)

The ASSET approach draws on the tools and techniques of human performance assessment, instructional science, and phenomenology to establish a multidimensional, interdisciplinary perspective to performing TEEs. This approach increases the breadth of the evaluation 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. A condensed version of the ASSET approach is illustrated in Figure 2.

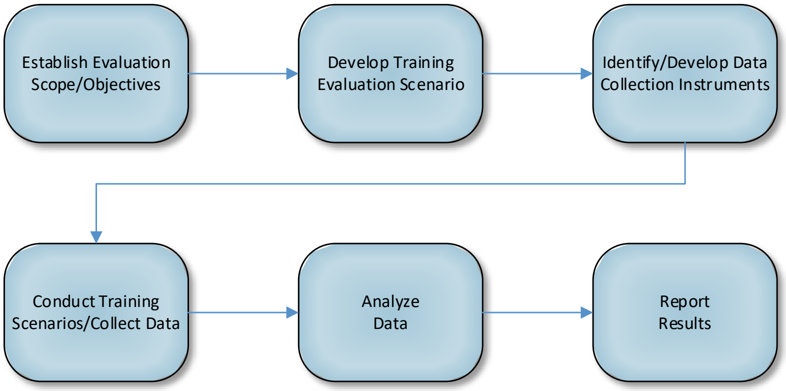
Establish Evaluation Scope/Objectives
Develop Training Evaluation Scenario
Identify/Develop Data Collection Instruments
Conduct Training Scenarios/Collect Data
Analyze Data
Report Results

Figure 2. ASSET Evaluation Approach

The ASSET approach is described in the following sections in the context of a use case to evaluate Army Aviation training. The U.S. Army Aviation Combined Arms Training Strategy (2016) emphasizes the use of Training Aids, Devices, Simulations, and Simulators (TADSS) to prepare Army aviation forces for future combat. This strategy highlights multiple types of environments that encompass a wide range of fidelity and cost. Some broad examples include game-based systems, moderate-fidelity trainers, and high-fidelity flight simulators. Of these environments, there is a high level of interest in the training potential of game-based systems. However, the effectiveness of game-based simulations requires further investigation (Sotomayor & Proctor, 2009; Whitney, Tempby, & Stephens, 2014). In particular, the use of game-based training as an adjunct to traditional simulation-based training has not been adequately evaluated.

A TEE was performed using the ASSET approach to conduct evaluations of three simulated training environments to determine the potential of lower-cost, game-based simulations to augment higher-cost, traditional simulation-based training. The training environments evaluated in the study were the Aviation Combined Arms Tactical Trainer (AVCATT; the current U.S. Army Program of Record for aviation collective training), a moderate-fidelity training simulator that integrates augmented reality helmet mounted displays (HMDs) to blend the physical cockpit with the virtual environment; the Virtual Battlespace 3 (VBS3) low-fidelity, first-person, games-for-training system operated on a desktop computer with commercial-off-the-shelf (COTS) flight controllers; and Microsoft Flight Simulator (MSFS), a commercially available flight simulator game that provides a similar level of fidelity and operation as VBS3. An operational flight trainer (OFT), a full-motion FAA Level D flight simulator, served as a real-world analog and was used for evaluation of the training environments.

The ASSET approach began with an identification of the scope and objectives of the evaluation. This was an essential part of the process, as it established the parameters for performing the rest of the evaluation. For the present use case, it was determined that the primary objective was to determine how and where lower-cost game-based training could be used as an equally effective adjunct to higher-cost simulation-based training for a particular set of aviation collection mission training tasks. Based on this evaluation objective, the following three evaluation questions were identified to establish the scope of the TEE: 1) Are there differences among the three simulated training environments?; 2) Are there differences in a real-world analog environment (OFT) based on the preceding simulated training environment?; and 3) Are there differences in the degree to which each simulated training environment corresponds to the real-world analog environment (OFT)?

Once the scope and objectives were established, the training scenarios that formed the basis of the evaluation were developed. The training evaluation scenarios involved a flight of UH-60A/L Blackhawk helicopters engaged in a collective air assault mission and consisted of a set of operationally demanding tasks and cognitive decision-making points. Operational tasks focused on mission events that are part of standard operating procedures or explicit items covered in mission and crew briefings. Cognitive decision-making focused on the pilot's specific choices and reactions to changing conditions during the mission scenario. These tasks and decision points directly related to the ability of the investigated training environments to support their execution and were part of the mission performance rubrics for the study.

The next step was to identify, develop, and collect data using a set of specific measures and data collection instruments that supported the objectives of the evaluation. An interdisciplinary set of empirically validated measures that contribute to training effectiveness were used. These measures aligned within the disciplinary areas of psychology, physiology, and phenomenology.

Psychology

Psychological measures included a mission performance rubric and questionnaires. The mission performance rubric consisted of 12 individual tasks and 5 decision points. Questionnaires from the psychology discipline were used to record a variety of subjective measures related to immersion, presence, workload, stress, and simulator sickness.

An Immersive Tendencies Questionnaire (ITQ; Witmer & Singer, 1998), version 3.01, as revised by the Université du Québec en Outaouais Cyberpsychology Lab, was administered at the beginning of the experimental session. Immersive tendencies were scored across four subscales: Focus (paying attention to current tasks), Involvement (interacting with current tasks), Games (becoming engaged within a scenario), and Emotions (experiencing fear, excitement, or other feelings). A Presence Questionnaire (PQ; Witmer & Singer, 1998), version 3.0, as revised by the Université du Québec en Outaouais Cyberpsychology Lab, was administered at the completion of each experimental session. The PQ assessed the degree to which participants experienced presence in each of the simulated environments, as well as the intensity of this experience as influenced by seven individual factors (realism, possibility to act, possibility to examine, quality of interface, self-evaluation of performance, sounds, and haptic). A Simulator Sickness Questionnaire (SSQ: Kennedy, et. al., 1993) was used to assess the level of discomfort experienced by participants in each of the simulated environments. The SSQ consists of items related to symptoms of simulator and motion sickness (eyestrain, headache, dizziness, etc.), clustered into three factors: Oculomotor, Disorientation, and Nausea. The Dundee Stress State Questionnaire (DSSQ; Matthews, et. al., 2002) differentiates 11 primary state factors relating to affect, motivation, and cognition. These primary state factors support three broader second-order factors: engagement (qualities of interest, motivation, and energy), distress (feelings of confidence, tension, and control), and worry (levels of self-esteem, self-focus, and cognitive interference). A short version of the DSSQ was used in the described study (Matthews, Emo, & Funke, 2005). A pre-task questionnaire was administered at the beginning of the experimental session and a post-task questionnaire was administered at the completion of each experimental session. The NASA-Task Load Index (TLX; Hart & Staveland, 1988) was used to assess each participant's perceived workload during the performance of the mission scenarios. The TLX is composed of six subscales that measure workload across the dimensions of mental demand, physical demand, temporal demand, effort, frustration, and performance. A separate global workload score is computed as the unweighted averages of the six subscale scores. The TLX was administered at the completion of each experimental session.

Psychological measures provided important data related to training effectiveness that is often overlooked in traditional TEEs. Factors relating to immersion, presence, workload, stress, and simulator sickness all correspond to the ability of a simulated training environment to support the positive performance of training tasks. Performance measures may also provide indications of differences between training environments.

Physiology

Physiological measures consisted of electrocardiography (ECG) and galvanic skin response (GSR). Both of these measures were captured using a Procomp Infiniti system. ECG is a direct measure of cardiac activity and one of the most common physiological measures of workload and stress in response to task demands. ECG measures included Inter-Beat Interval (IBI), Heart Rate Variability (HRV) and Beats per Minute (BPM). Increases in BPM have been associated with increases in workload and this particular measure is more sensitive to physiological workload (Wilson & O'Donnell, 1988; Jorna, 1993). HRV is generally associated with cognitive workload rather than physiological workload. As such, it reflects engagement in effortful information processing (Jorna, 1993). Increases in cognitive workload of task demands are associated with decreases in HRV (an inverse relationship; Mulder, Waard, & Brookhuis, 2004).

GSR is a measure of emotional stress and nervous tension based on the electrical conductance of the skin (Mundell, Vielma, & Zaman, 2016). Increases in GSR are associated with increases in stress and tension (Shi, Choi, Ruiz, Chen, & Taib, 2007). GSR drift, the difference between the upper and lower levels of galvanic skin response, is a measure of emotional arousal related to stress. Absolute drift, in particular, is the absolute change in raw GSR from the beginning to the end of a session (Mundell, Vielma, & Zaman, 2016). Absolute drift reveals slow variations in the GSR signal. GSR Maximum Increase Drift is the absolute difference in raw GSR from the minimum point to the end of the session (Mundell, Vielma, & Zaman, 2016). Maximum increase drift gives a measure of trends in the GSR signal existing at the end of the session.

These measures captured the direct, real-time physiological responses of study participant's as they were engaged in mission scenarios within the simulated training environments investigated in this study. This provided an additional dimension of training effectiveness that helped broaden the evaluation effort.

Phenomenology

Study participants were interviewed at the end of each experimental session to collect first-person experiential data for each simulated training environment. The interview method was based on Petitmengin (2006) and implemented following the guidance provided by Bockelman, Reinerman-Jones, and Gallagher (2013). Participant interviews consisted of questions designed to focus the participant's attention on the real-time subjective experience of performing the mission scenario in a particular simulation environment. Questions such as "Describe what it is like performing the mission in the [*type of simulator*] environment." and "Tell me your thoughts as you progress through the mission." provided opportunities for participants to relate their direct experiences with the simulated environments. Copilots were also interviewed after each experimental session. Although they were confederates in the study, data collected from copilot interviews provided an additional source of evaluation information. These interviews helped capture the ability of the simulated environments to support mission training tasks in terms of graphics, controls, responsiveness to inputs, and representation of flight and mission characteristics.

Summary

Evaluating the effectiveness of training in simulation domains cannot be adequately accomplished by standard TEE approaches and methods. The ASSET approach represents a novel method for conducting TEEs in simulation training environments that transcends the limitations of standard approaches. ASSET is based on the procedures and rigor of a research methodology, but is specifically optimized to conduct TEEs for simulation training. Its interdisciplinary focus on human performance assessment, instructional science, and phenomenology increases the scope of the evaluation effort to more fully capture the range of factors that contribute to training effectiveness in dynamic, interactive simulation training environments. Beyond its application in the described use case, the ASSET approach provides a powerful methodology for evaluating simulation training in any context. Its use becomes essential when the objective of the evaluation extends beyond a determination of the value of a training program and into the need to inform decisions concerning how, what, when, and where simulation training will be implemented to meet specific training requirements.

Acknowledgements

This paper was prepared in conjunction with work sponsored by the U.S. Army Program Executive Office for Simulation, Training, and Instrumentation (CA# W911NF-14-2-001). The U.S. Army Program Executive Office for Simulation, Training, and Instrumentation nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product, or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights. The views expressed in this paper are not necessarily those of the U.S. Army Program Executive Office for Simulation, Training, and Instrumentation.

References

Army Aviation Training Strategy. (January 2016). U.S. Army Aviation Center of Excellence. Fort Rucker, AL.

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

Bockelman P, Reinerman-Jones L and Gallagher S (2013) Methodological lessons in neurophenomenology: Review of a baseline study and recommendations for research approaches. *Frontiers in Human Neuroscience,* **7**:608.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) *Human Mental Workload.* Amsterdam: North Holland Press.

Jorna, P.G.A.M. (1993). Heart-rate and workload variations in actual and simulated flight. *Ergonomics, 36*(9), 1043–1054.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, *3*(3), 203-220.

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.

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., et al. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, 2, 315–340.

Matthews, G., Emo, A. K., & Funke, G. J. (2005). *A short version of the Dundee Stress State Questionnaire*. Paper presented at the Twelfth Meeting of the International Society for the Study of Individual Differences, Adelaide, Australia.

Mulder, L. J. M., de Waard, D., & Brookhuis, K. A. (2004). Estimating mental effort using heart rate and heart rate variability. In N. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. Hendrick (Eds.), *Handbook of Human Factors and Ergonomics Methods* (pp. 201-208). Boca Raton, FL: CRC Press.

Mundell, C., Vielma, J. P., & Zaman, T. (2016). *Predicting performance under stressful conditions using galvanic skin response*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1606/1606.01836.pdf.

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive Sciences*,*5*(3-4), 229-269.

Shi, Y., Choi, E. H. C., Ruiz, N., Chen, F., & Taib, R. (2007). Galvanic skin response (GSR) as an index of cognitive workload. In *ACM CHI Conference Work-in-progress*.

Sotomayor, T., & Proctor, M. (2009). Assessing Combat Medic Knowledge and Transfer Effects Resulting from Alternative Training Treatments. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 121-134.

Whitney, S., Tempby, P., & Stephens, A. (2014). A Review of the Effcetiveness of Game-based Training for Dismounted Soldiers. *Journal of defenese Modeling and Simulation*, 319-328.

Wilson, G. F., & O'Donnell, R. D. (1988). Measurement of operator workload with the neuropsychological workload battery test. *Advances in Psychology, 52*, 63-100.

Witmer, B.G. & Singer. M.J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, *7*(3), 225-240.