Considerations for training evaluations of emerging technologies

Julian Abich IV, Jennifer Murphy, Morgan Eudy
Quantum Improvements Consulting
Orlando, FL
jabich@quantumimprovements.net,
jmurphy@quantumimprovements.net
meudy@quantumimprovements.net

John P. Killilea
Naval Air Warfare Center, Training Systems Division
Orlando, FL
john.killilea@navy.mil

ABSTRACT

Across many domains, new training systems are often acquired and implemented without first determining the appropriateness of these systems in terms of achieving specific learning objectives and training outcomes. Further, evaluations of their training effectiveness and efficiency are either not conducted or are based on methodologies that are not suitable for extracting proper recommendations or informing policy changes. As the state-of-technology makes available new training systems that seem viable as cost-effective training solutions, best practices for implementation must be considered beforehand. On the surface, emerging technologies, such as augmented reality (AR), virtual reality (VR), and mobile platforms may appear as effective and efficient solutions to almost any type of training. Yet, rather than acquire these systems based solely on face validity, adopting a learning science approach will result in training systems that bestow greater benefit to the trainee and lead to a larger return on investment. This paper will take a theoretical approach backed by empirical evidence extracted from the literature to derive considerations for training with emerging technologies, with an emphasis on AR, VR, and mobile platforms. A methodology will be presented that lays out the approach for an experiment to evaluate the effectiveness of a cross-platform training system. The topic presented will generate theoretical discussion of current training effectiveness models and frameworks and suggest the need for a new approach to guide effectiveness evaluations throughout the training process. Practical results of this preliminary evaluation will generate guidance for researchers and acquisition personnel when conducting training effectiveness evaluations to inform procurement decision-making.

ABOUT THE AUTHORS

Julian Abich IV is a Senior Human Factors Engineer at Quantum Improvements Consulting, LLC. He has over 10 years focused on applying human factors, ergonomics and modeling & simulation principles to the assessment, prediction, and improvement of human performance. His current research efforts focus on the application of human performance data from complex training systems for improving training efficiency and effectiveness and enhancing human-system interaction. He holds a Ph.D. in Modeling and Simulation with a specialization in Human Factors from the University of Central Florida.

Jennifer Murphy is the CEO of Quantum Improvements Consulting, LLC. She has over 14 years of military selection and training research experience, with an emphasis on leveraging innovative technologies for improving training in a measurably effective way. Her current research focuses on developing assessments of Warfighter performance to enable adaptive training, predictive modeling, and improved training effectiveness. She holds a PhD in Cognitive and Experimental Psychology from the University of Georgia.

Morgan Eudy is a Human Factors Intern at Quantum Improvements Consulting, LLC. He received his M.S. in Aviation Human Factors from the Florida Institute of Technology in 2018. As an intern with Quantum Improvements Consulting, Morgan acquired experience implementing technology-based training systems, performing training effectiveness analysis, and collecting human subjects research in support of a variety of military projects. His research interests are in human performance, emerging technology, and performance support tools.

John P. Killilea is a Research Psychologist in the Basic and Applied Training and Technology for Learning and Evaluation (BATTLE) Lab at the Naval Air Warfare Center, Training Systems Division (NAWCTSD). The mission of the BATTLE lab is to conduct and manage science and technology (S&T), research and development (R&D), transition and acquisition consultation efforts through the application of cognitive science, behavioral research and training evaluations to improve training and human performance in a variety of learning environments. John graduated with a Ph.D. in Modeling and Simulation from the University of Central Florida with a specific emphasis on training. Prior to working at NAWCTSD, he worked as a research psychologist supporting the Army Research Institute.

Considerations for training evaluations of emerging technologies

Julian Abich IV, Jennifer Murphy, Morgan Eudy
Quantum Improvements Consulting
Orlando, FL
jabich@quantumimprovements.net,
jmurphy@quantumimprovements.net
meudy@quantumimprovements.net

John P. Killilea
Naval Air Warfare Center, Training Systems Division
Orlando, FL
john.killilea@navy.mil

INTRODUCTION

All too often technological solutions are implemented into training programs without proper evaluation of their training impact. The appeal of commercially available, low cost, distributed, customizable solutions tends to overshadow the true benefits for training and therefore, overvalue the perceived return on investment. It is imperative that proposed technological solutions be vetted through empirical approaches to derive data-driven results for acquisition decision-making. Otherwise, resources allocated to training system procurement will be wasted and more importantly, training insufficiencies will result. Hence, an examination of the current training effectiveness practices must be conducted in order to identify inference limitations and recommend a more robust approach that will result in evaluations to capture the true training benefits.

The goal for this paper is to set the groundwork for the standardization of training effectiveness evaluations by presenting key considerations that should be addressed. A practical review of previous research will identify examples that portray opportunities for improvements of current training effectiveness evaluations that have been conducted across domains focused on augmented reality (AR), virtual reality (VR), and mobile applications. Lastly, a proposed training effectiveness evaluation design will be described to illustrate a prescribed approach.

Training Effectiveness Evaluations

Training effectiveness is a pervasive term that spans across domains yet yields so much misunderstanding when implemented in both the laboratory and field. Tamkin, Yarnell, and Kerrin (2002) eloquently described training evaluation as "a bit like eating five portions of fruit and vegetables a day' everyone knows that they are supposed to do it, everyone says they are planning to do better in the future, and few people admit to having got it right." Efforts have been made over that past few decades to provide rationalizations and guidelines for conducting training effectiveness evaluations (Boldovici, Bessemer, and Bolton, 2002). They warn against making rationalizations for "junk training evaluations (Boldovici et al., 2002)," referencing Cohn (1994) who described "junk science as a modifier for any evaluations that permit no valid inferences about training effects." Further, it has been argued that failing to take the necessary precautions to ensure stated benefits truly exist after an evaluation is a violation of professional ethics (Bates, 2004; Beauchamp & Childress, 1983). Therefore, to increase the inferencing validity of training effects, proper evaluations should base their approaches on training models and frameworks. Unfortunately, many of them are lacking the guidance and specificity needed for evaluating complex training systems, such as those integrating AR and VR.

The current training model standard is Kirkpatrick's Four-levels (Kirkpatrick, 1976), including the New World Model (Kirkpatrick & Kirkpatrick, 2016), which laid the foundation for many other models (Passmore & Velez, 2012). This model has been the subject of extensive scrutiny, stating that it is more of a taxonomy, lacks the ability to address the effectiveness of training, and is not supported by empirical evidence (Bates, 2004; Holton, 1996; Tamkin, Yarnall, & Kerrin, 2002). Further, the hierarchical structure of these models assumes that training success is dependent on achieving each sequential level, but there may not always be linear relationships among the various factors that contribute to training (Mathieu, Tannenbaum, & Salas, 1992). Other models, such as the Learning Transfer System Inventory (LTSI) (Bates, Holton, & Hatala, 2012) and Integrated Model of Training Evaluation and Effectiveness (IMTEE) (Alvarez, Salas, & Garofano, 2004) leveraged the concepts of Kirkpatrick's model, but oriented their approaches to identify the relations among contributing factors to training effectiveness. Hence, these types of models should guide researchers when designing training effectiveness evaluations that will capture the data needed to make empirically-based decisions for training system acquisition.

As technological systems advance and emerge, so to will assumptions that these new systems can provide benefit to any type of training. There is a tendency to assume that with new technology a positive correlation exists between the number of features and characteristics of a training system and the benefit it provides, rather than realize it is how the systems are implemented (Salas, Bowers, & Rhodenzier, 1998). Although researchers should exercise due diligence when evaluating all training systems, the importance of adhering to a structured scientific approach is essential as the number of potential technological solutions becomes more accessible. The argument for stringent training effectiveness evaluations is not novel, but the criticality of addressing this issue at the cusp of a technological disruption will help guide appropriate applications of this technology for training.

Reviewing the AR and VR literature for published training effectiveness evaluations revealed that this body of research is still in its infancy. Further, most of these evaluations focused on the development of the technology itself, rather than on the benefits afforded by this technology for training. The likely reason for this is that AR and VR technology has only matured over the last decade to offer affordable consumer products that could be evaluated in the training realm. Further, development kits have recently become available to allow consumers to create their own content and experiences on personal devices, such as tablets, and on commercial hardware platforms, such as the HTC Vive. Meaning, not enough time has been available for designing and developing training content, and evaluating the application in these new technological systems, especially their long-term impacts. Of the AR research found, these types of training systems were used as a learning tool for users to gain the knowledge, skills, and abilities that would later be practiced to further enhance mastery, such as knowledge in physics (Lin, Duh, Li, Wang, & Tsai, 2013), spatial knowledge of historical locations (Chang, Hou, Sung, Chang, 2015), quality assurance in assembly task (Webel et al., 2013), and battle planning distance estimation (Schmidt-Daly, Riley, Hale, Yacht, & Hart, 2016). The VR literature was even more scarce, but what can be gathered is that these types of training systems were used as a practice environment for tasks such as welding (Stone, Watts, & Zhong, 2011), military corrosion protection training, (Webster, 2014), and spatial navigation (Stroud, Harm, & Klaus, 2005). Although, AR and VR gained their foothold in the entertainment and gaming industry, they are now transitioning to more practical applications. With this, it is important that practical applications are designed based on a foundation of learning theories, human factors, and cognitive science. Therefore, it is expected that the next decade will see an exponential growth of AR and VR training effectiveness evaluations across a variety of domains.

On the other hand, the mobile learning literature has been established much longer than both AR and VR training systems. Electronic learning (e-learning) gained momentum with the affordability of personal computers and transitioned further when personal computing became more mobile and demand for learning outside of the classroom increased. Although early efforts of e-learning tended to simply digitize course material and textbooks (Lindner, 2007), today, learning theories and frameworks have been adapted and developed to guide mobile learning content and analysis (MacCallum & Parsons, 2016). In educational settings, mobile learning tends to be blended with traditional classroom-based education, providing students with opportunities to learn outside of the classroom (Haythornthwaite, Andrees, Fransman, & Meyers, 2016). Mobile learning tends to be suited for learning on-the-go and leverages microlearning, which is presenting the user with "very small preconfigured information objects (Lindner, 2007)." Mobile learning provides unique opportunities for training and therefore needs further evaluation for practical applications.

A full review of this literature was outside the scope of this effort but select papers were used to discuss opportunities for improving the way in which training effectiveness evaluations should be conducted. From this, considerations were drawn from exemplar cases and inconsistencies and misconceptions present in these evaluations. The purpose here is to learn from previous research by taking a constructive approach to develop a list of considerations that will help future evaluations of training systems, the focus here on AR, VR, and mobile training systems. The first step is to define what is meant by AR, VR, and mobile to frame the discussion for training.

Augmented reality

AR is a form of extended reality that is the middle ground between the real world and full virtual reality. Through AR, "things," such as objects, people, information, etc., are superimposed on items in the real world (Figure 1). Inversely, AR also can remove items from the real world, or at least make it appear that way. A key distinguishing feature of AR is that it allows the user to interact with both virtual elements and the real-world simultaneously. Further, AR is not restricted to just visualizations, as haptic feedback, smells, and sounds can all be added to the real world. Since there are many types of AR technologies and it can be described in many ways depending on the type of technology that is being used, to provide an encompassing definition Azuma's (1997) early survey of AR stated it must adhere to three characteristics: 1) combines real and virtual, 2) is interactive in real-time, and 3) is registered in three dimensions.

Figure 1. Example of AR.

Virtual reality

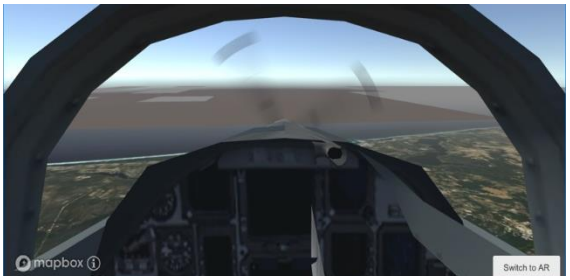

Figure 2. Example of VR. POV from the cockpit of an airplane.

VR is another form of extended reality comprised of a system interface that users employ to access virtual worlds (VWs) or virtual environments (VEs). VR utilizes specialty hardware that increases the immersion, or sensation of being in a virtual space for users (Loomis, Blaskovich, & Beall, 1999). The two most common virtual reality systems use head tracking via a head-mounted display (HMD), or they use a computer automatic virtual environment (CAVE) system (Coomans & Timmermans, 1997). Both VR platforms substantially increase the user's perception of immersion compared to a traditional 2-D computer monitor or large-screen TV. Furthermore, a VR system provides a first-person point of view (POV) (Figure 2). Many VR systems are multi-modal, utilizing haptic gloves or spatial audio, and may add to participants' level of immersion, but are not necessary requirements. VR differentiates itself from VE by requiring specialized hardware that increases user immersion and replaces the character with a first-person POV. This increases a participant's sense that they are the operative agent in the virtual space (Kalawsky, 1993).

Mobile

This category may not be as clearly defined as AR and VR, because the concept of mobile-based platforms could potentially encompass the other two. Mobile learning refers to "learning that happens across locations, or that takes advantage of learning opportunities offered by portable technologies (Sharples, 2009)." The portable technologies that are the focus here are smartphones and tablets (Figure 3). For the most part, smartphones are mobile phones with many functions of a computer, such as internet access, operating systems capable of downloading applications, touchscreens interface and voice recognition. They offer other functionalities beyond desktop and laptop computers that include motion-based sensors such as accelerometers, gyroscopes, and magnetometers. Tablets offer the same functionality as smartphones, but usually with a larger screen size, which provides more screen real-estate.

Figure 3. Example of mobile learning application.

Defining these technologies and their requirements helps to frame an understanding for their applications. Although there are overlapping features of these technologies, their practical applications will differ since AR allows interaction with the real-world simultaneously, but VR encompasses the user in a fully synthetic environment that may or may not reflect the real-world. AR technology is commonly used through mobile devices utilizing location- and recognition-based functions. VR can also be used through mobile devices with the use of specialty head-mounted hardware, such a Google cardboard and Gear VR. Thus, mobile technology seems to be most flexible platform, but the training benefits of each application may differ. By evaluating the features and functions of each platform, distinct applications of each becomes discernible.

TRAINING EVALUATION CONSIDERATIONS

Although the training effectiveness models and theories provide the conceptual approach, there are practical considerations that need to be addressed in order to the collect quality evaluation data. The considerations below were derived from the review of AR, VR, and mobile training effectiveness and learning literature. Most of these considerations may seem logical, which they are, but the question then is, "Why are these considerations neglected time and time again?" Most often there are resource limitations, such as access to end-users, time, and personnel, that preclude appropriate evaluations. Yet, in other instances, training effectiveness evaluations are conducted after the systems have been employed which increases the chances of training with ineffective systems. Regardless of the reasons, adjustments could be made to evaluation designs early on to deal with resource limitations without sacrificing sound evaluation processes. The list identified below is not exhaustive, but by addressing each either during the design, execution, or analysis, training effectiveness evaluation results will provide data from which confident inferences can be made.

Identify the Problem and Research Questions

The impetus for conducting training evaluations should stem from an identified gap or issue that needs to be rectified. Conducting training effectiveness evaluations is only as good as its intended purpose (Russ-Eft & Preskill, 2009). Working with stakeholders and customers, the training needs and research questions that must to be addressed should be clearly defined. The answers to these research questions should provide information that will help make the determination if the training system effectively fulfills the need. This may seem like a logical first step, but often training system developers fail to address targeted needs and instead try to find training gaps to fill after development. Even more detrimental to training is when systems are implemented based on novelty and face validity. Not only is this a waste of resources, but there are safety concerns that may result from improper training support.

Consideration 1: Just because new training systems exist, does not mean they will be effective. Make sure the training system is designed to address a training need and that the evaluation focuses on collecting data to answer the research questions before the systems are fully employed.

Identify Learning Objectives and Training Outcomes

All training systems should guide users to gain a new knowledge, skill, or acquire new information that will be used on the job. Therefore, training systems should guide users to achieve learning objectives and/or training outcomes. There is no guarantee that achieving learning objectives or training outcomes will transfer to the real-world, but if training systems are designed to address a training gap, the likelihood of transfer improves. Often, there are programs of instruction (POIs) that identify the training expectations. Working with stakeholders and customers, learning objectives and training outcomes can be extracted or updated with the support of subject matter experts (SMEs) to ensure they still meet the training goals. Furthers, learning and cognitive theories are important to understand as they support the expectations of the learning objectives and training outcomes (MacCallum & Parson, 2016).

Consideration 2: Learning objectives and training outcomes provide users with the purpose for the training. Make sure to clearly defined learning objectives and training outcomes, otherwise the training will likely fail to meet the training needs.

Understand the Training Systems Under Evaluation

It is important to understand features and functions of the training systems under evaluation in order to help determine their most appropriate applications. AR, VR, and mobile platforms offer unique features and functions that can be leveraged in different ways to meet the training needs. Knowing if a system captures motion data or allows a user to simulate a real-world scenario can help determine which learning objectives and training outcomes would be met by those platforms and how best to train to achieve those goals. Issenberg (2005) generated a list of features and uses of high-fidelity simulators that were found to lead to effective learning that can be leveraged across training contexts, but fidelity is not always a determinant of effective training systems (Salas, Bowers, & Rhodenizer, 1998). Hence, understanding how and when to use the system may be more important than the capabilities of the system.

As with almost any new technology, the novelty of these systems may mediate significant results during evaluations (Clark, 1989; Clark & Sugrue, 1988). In a meta-analysis examining the effects of VR-based instruction on student's learning outcomes, novelty effects were found to significantly impact learning outcome gains (Merchant, Goetz, Cifuentes, Keeny-Kennicutt, & Davis, 2014). The novelty effect will have a higher likelihood of occurrence with a training system that provides immersive, 3-D visualizations of training content.

Consideration 3: Knowing the features and functionalities of the platforms will allow researchers to better align training systems with learning objectives and training outcomes across the training stages. Be aware that the novelty of the training system may be influencing the training effectiveness evaluations and therefore, precautions should be taken when designing the evaluation and analyzing the data.

Clearly defined measures and metrics

Referencing the learning objectives and training outcomes, measure and metrics should be defined and identified in order to capture the required data for the evaluation. The metrics are often derived from conversations with the stakeholders and customers. The measures used to gather the data should be validated, regardless if they are designed specifically for the evaluation. Using the POI, established training material, and SME interview responses, measures can be developed, then vetted through stakeholder, customer, and SME validation to ensure the measures are capturing the appropriate data. Additionally, the quality of the data is just as important because the outcome of the statistical analyses is only as good as the data used.

Consideration 4: The metrics that are determined to be relevant should have associated measures to collect the appropriate data to answer the research questions or support the stated hypotheses. Limited resources are always a concern when executing training effectiveness evaluations, therefore make sure to not only identify all the independent and dependent variables, but also the modifying and mediating variables. Collecting data outside of this scope should be avoided.

Sample Population

Sample size tends to be the most widely neglected consideration when conducting training effectiveness evaluations, because it is difficult to control. Without a large enough sample size, valid inferences cannot be made from the observations, and conclusions drawn may be fallible. Although power analyses should be conducted to determine the appropriate number of observations needed to confidently detect significant differences before conducting an evaluation, factors, such as sample attrition, will likely reduce the sample size. Therefore, a power analysis should be conducted following the evaluation to indicate the power based on the collected observations. Additionally, it should not be assumed all participants have a similar background or experience, even if they from a similar population (e.g Tang et al., 2003). Participants bring a variety of individual differences that can influence their training outcomes. By identifying potential individual differences, measures can be administered to capture the data during evaluation.

Consideration 5: First, make sure your sample size is large enough to make inferences based on the data collected. Second, make sure you learn about your sample and the individual differences they bring to the training, as this will likely impact the results. Third, make sure measures are in place to capture individual difference data that could help with data interpretation (usually before any conditions are run with demographics questionnaires).

Experiment Design

The design of the experiment greatly impacts the results of the evaluation (Campbell & Stanley, 1963; Cook & Campbell, 1979). There are many threats to research design that have been overlooked in many evaluations, and the inferences drawn from those data sets are flawed. Without going into great detail about each threat, the two commonly identified experiment design errors were lack of counterbalancing and recognizing the impact of order effects. Within a repeated-measures design, the training in one scenario will impact the training in another scenario (e.g. Adams, Klowden, & Hannaford, 2001), therefore reducing the chance of drawing distinct conclusions from any evaluation results. By counterbalancing the exposure to each condition, order effects can be controlled or excluded from interfering with the data set. Between-subjects designs will help mitigate this threat but will require more participants for the evaluation and thus potentially introduce more individual differences.

Consideration 6: Well thought-out research designs will help ensure that the data collected informs the decisions regarding the effectiveness of the training systems. Make sure to randomize participants in each group so that each has equal chance to participant in any condition. If using repeated measures, make sure to have enough participants for each experiment condition order to reduce the chances of order effects intruding your results.

Training Transfer

Methods for measuring transfer of training have been established (Lathan, Tracey, Sebrechts, Clawson, & Higgins, 2002). Do not assume transfer from simulation-to-simulation will reflect actual performance in real-world (Bell & Waag, 1998). Another important aspect to validly capture training transfer is the time delay between training and real-world application (e.g. Gavish et al, 2015). This step is by far the most difficult to evaluate as this requires evaluation over time which may not be possible. Additionally, it is also difficult to parse out the contributions that the training had because of other factors that may contribute to real-world success outside of the training. Attempts are being made through various approaches to capture data from trainees and professionals throughout their career to determine the effectiveness of training and identify areas for improvements to provide customizable remediation.

Consideration 7: Even through measuring transfer of training to the real-world may be difficult, it should always be the desired outcome of training. If it is possible to evaluate the transfer, make sure this is done at multiple instances over time following the training to not only determine if there was a transfer, but how long the training remained effective.

PROPOSED EXPERIMENT DESIGN

Taking into account the training effectiveness evaluation considerations, an evaluation is proposed below. The purpose for this evaluation is to take an empirical approach for evaluate three training platforms to determine the appropriateness of utilizing each platform to support different aspects the training process. The context identified by the government sponsor was aviation course rules training. The following section describes a proposed approached to conduct this evaluation, illustrating the application of the training effectiveness evaluation considerations described above.

Use Case

The use case identified by the government sponsor was Naval aviation course rules training for entry-level Naval flight students (NFSs). Course rules refers to the policies and procedures in place for an airport and the surrounding airspace. These rules comprise five main areas: 1) airfield and general info, 2) ground operations, 3) takeoff and departures, 4) home field entry, and 5) landing and parking. The topic was scoped down to focus on home field entry.

A front-end analysis was conducted to understand the course rules training procedures. Chief of Naval Air Training (CNATRA) issued documents were reviewed as the primary source of currently validated information. These documents covered the procedures NFSs are expected to master, but they do not indicate how it is all taught. Therefore, an interview was conducted with a SME that was both a flight instructor for entry-level NFSs and a pilot advancing his career as a jet fighter. This granted a unique opportunity to generate an understanding of the current training process from two different perspectives. The interview provided valuable information unattainable from training procedural manuals that included classroom and in-flight training strategies, training evaluation criteria, and challenges that students and instructors face during training,

Although the U.S. Navy is modernizing their training program under the Sailor 2025 initiative, course rules training is still taught in a traditional manner. As part of the ground school, it is taught in a classroom through face-to-face instruction with course rules materials consisting of a procedural flight manual documents and PowerPoint slide deck. Students are not required to have prior knowledge to successfully complete course rules training but are encouraged to familiarize themselves with the content prior to ground school. An issue raised is that there is no teaching standard for instructors to follow and as a result, the training quality can often be dependent on the experiences of the instructors.

Three major challenges that NFS students face were identified from the interview. First, students have difficulty gaining an understanding of where they are in time and space. This is a two-sided issue for both students and instructors. In the current training format, students struggle with gauging how long it should take to get from one point

to another without having any aids that accurately and realistically portray speed. Congruency of materials to accurately reflect the ways in which they are presented in the real-world can moderate multisensory learning (Shams & Seitz, 2008). Instructors face the issue of trying to convey this type of information with 2-D static images. The second challenge is that student's have difficulty memorizing the vast amount of procedural information. Although they do have access to an in-flight guide on their kneeboard, the goal is for students to memorize this information to the greatest extent and only use the kneeboard as a last reference. The third challenge for students is developing in-flight situation awareness. Students tend to struggle with knowing their geographical location and altitude, which becomes critical when determining the appropriate runway approach.

Methodology

Participants

The ideal approach is to use a sample of entry-level NFSs. If access cannot be granted to this population, then using a general population, such as university students, may still provide valid results since no pre-existing knowledge is expected for course rules training success. There may be factors that enable successful training, such as interest in or experience with aviation, and it may be appropriate to find participants with this background.

Measures

Pre-task assessments: Pre-task measures are necessary to gather information that could account for individual differences in training. Demographics questionnaires will be developed to capture context relevant information, such as experience in a particular domain. These assessments will also account for any enabling factors that may contribute to the success of the training, such as various cognitive or personality factors (Cannon-Bowers, Salas, Tannenbaum & Mathieu, 1995).

Performance: Pausing the training to administers an assessment interrupts the training process, potentially influencing the results of the study. The proposed approach will be to design assessments that are embedded into the training. Embedded assessment is "the process of measuring knowledge and ability as part of a learning activity rather than after the fact, when it is only an approximation of learner behavior (Underwood, Kruse, & Jakl, 2010)." Course rules involves communicating specific information at designated points in time. One way to measure training is to evaluate if a user communicates at the appropriate time, and if that information is accurate.

Usability: The application of all technology requires user feedback regarding the satisfaction with using the system, the effectiveness of the system to help users achieve a goal accurately, and the efficiency through which task goals were achieved. Data can be captured both during the users' interaction with the system through observations and system input and with post-task questionnaires and debrief interviews.

Transfer of training: Training should ultimately transition to the real-world. The difficulty in capturing this type of data is isolating the impact of the training from other factors that can contribute to task performance. Further, this usually involves long-term evaluation which is often not possible to conduct due to resource or security limitations. Attempts may be made to capture how well course rules training is executed by comparing students that use the training system to either the classes that have gone through the program in the past and/or after if they do not implement the training system. Since the students are evaluated by course instructors in-flight, their feedback would be most telling of transfer success.

Apparatus/hardware

AR: The platform that will be used to display the AR training content will be a tablet (e.g. Samsung 7" Galaxy) in combination with course rules material. Tablets will be used because for this context, all the interactions will be done through the AR interface. Course rules materials, such as military and sectional maps, will be embedded with AR markers to trigger the AR visualizations. The tablet's functionality, such as the accelerometer and gyroscope will be accessed to facilitate the AR interaction.

VR: The HTC Vive will be used to display the VR content. The resolution offers 1080 x 1200 pixels per eye with a refresh rate of 90 Hz and field of view 110 degrees. Through the tracking system, boundaries are visually present if users near tracking limitations for safety. The microphone will also be utilized for accepting voice input from the user.

Mobile: Although the mobile content could also be displayed through a smartphone, which is how it being designed, the training content will also be displayed through the tablet. The difference will be that all of the content will be local to the tablet and not require external markers. Similar to the AR system, the mobile platform will also take advantage of the accelerometer and gyroscope functionality. Again, the microphone will also be utilized for accepting voice input from the user.

Research Design/Procedure

A within-subjects design will be used to evaluate the benefit each platform provides the student throughout the training process. Based on the features and functions available for each platform and the way in which these systems have been implemented in past research, the order in which students will interact with the systems will be AR, VR, then mobile. The AR system is expected to provide students with foundational information of course rules. This will be the first-time students are exposed to the content. The system will not inundate the students with an overwhelming amount of required interaction but provide them with visualizations that reflect the real-world and procedural information needed to successfully complete course rules. After completing their training on the AR platform, they will then move on to the VR. The VR platform is envisioned to be an opportunity for the students to apply their training to a simulated real-world scenario. The immersive environment of the VR will allow students to execute behaviors as they would in the actual aircraft, allowing them to develop the knowledge and skills they need to transfer to the real-world. The mobile application is seen to maintain their training and provide a just-in-time refresher. This platform will offer students interactive opportunities to stay up to date with their training between flight times by enabling mobile learning.

Statistical Analysis

Most training effectiveness evaluations that compare a traditional training system to a modification of that system or a new one altogether usually implement comparative statistical analyses, such as *t*-tests and ANOVAs. The goal here is to determine the relationship among factors and behavioral interactions in one system and performance in another. It would not be appropriate to compare the AR, VR, and mobile platforms, because the main research question is not interested in determining if one system is better than the other for overall aviation course rules training, but rather, how does interacting with one system impact the performance in another. Through regression analyses, one variable (or several) can be used to predict another with some level of confidence. For instance, a research interest may be determining whether the length of time students spend interacting with a module in the AR platform can predict how accurately they will perform a task in VR. Conducting a direct comparison between these platforms may set up "straw-man" evaluations when it is clear that one system may be more beneficial than another to support a training task or stage in the process. Executing that type of evaluation may violate professional ethical principles of beneficence (Bates, 2004), hence, adhering to the approach proposed above. This approach allows all the possible variables of interest to be tested as a fit for the predictive (i.e. regression) model. This predictive capability of the empirical evidence is powerful when determining the suitable implementation of each type of training platform during the training process.

CONCLUSION

The goal for this paper was to emphasize the need for stringent training effectiveness evaluation approaches. The training systems of interest here were based on AR, VR, and mobile platforms. Each platform offers unique features and functions that can be utilized to meet specific learning objectives and training outcomes during the training process. Based on the training and learning literature that implemented these platforms, considerations for structured training effectiveness evaluations were extracted. These considerations tend to be issues that continuously appear in the training effectiveness literature, especially regarding what has been published with the AR, VR, and mobile training systems. As the technology for these training systems advances and new ones emerge, the necessity for structured standardized training effectiveness evaluations will need to be established.

Following, the identification of training evaluation considerations, an experiment was proposed that illustrated the approach to implement and account for the training effectiveness considerations. The design is focused on determining the appropriateness of implemented training technology throughout the training process and identifying how interaction with one system will predict the outcome in another system. It may be that current training frameworks lack the ability to account for the complexities of integrating new technological products into training solutions. Therefore, there is a need for a re-evaluation in terms of addressing the trainee across all steps of a training process and evaluate when certain training platforms will provide the greatest benefit. The results of the experiment will help

inform whether the current training models and frameworks account for the complex training solutions offered by new training systems, or if a new approach must be established.

ACKNOWLEDGEMENTS

This research was sponsored by NAWCTSD and was accomplished under Contract No N68335-19-C-0089. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of NAWCTSD or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

Adams, R. J., Klowden, D., & Hannaford, B. (2001). Virtual training for a manual assembly task. *Haptics-e, 2*(2), 1-7.

Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review, 3*(4), 385-416.

Azuma, R.T. (1997). A survey of augmented reality. *Presence, 6*(4), 355-385.

Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning, 27*(3), 341-347

Bates, R., Holton III, E. F., & Hatala, J. P. (2012). A revised learning transfer system inventory: factorial replication and validation. Human Resource Development International 15(5), 549–569.

Bell, H.H. & Waag, W.L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology, 8*(3), 223-242.

Beauchamp, T. L., & Childress, J. F. (1983). *Principles of biomedical ethics (2nd ed)*. New York: Oxford University Press.

Boldovici, J.A., Bessemer, D.W. & Bolton, A.E. (2002). *The elements of training evaluation*. Alexandria, VA: U.S. Army Research Institute.

Campbell, D. & Stanley, J. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand-McNally.

Cannon-Bowers, J.A., Salas, E., Tannenbaum, S.I., & Mathieu, J.E. (1995). Toward theoretically based principles of training effectiveness: A model and initial empirical investigation. *Military Psychology, 7*(3), 141-164.

Chang, Y.-L., Hou, H.-T., Pan, C.-Y., Sung, Y.-T., & Chang, K.-E. (2015). Apply an augmented reality in a mobile guidance to increase sense of place for heritage places. *Educational Technology and Society, 18*(2),

Clark, R. (1989). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445–459.

Clark, R. E., & Sugrue, B. M. (1988). Research on instructional media 1978-88. In D. Ely (Ed.), *Educational Media and Technology Yearbook*. Englewood, CO: Libraries Unlimited, Inc.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.

Coomans, M. K. D., & Timmermans, H. J. P. (1997). Towards a taxonomy of virtual reality user interfaces. *Proceedings of 1997 IEEE Conference on Information Visualization (Cat. No.97TB100165)*, (August 1997), 0 – 5. doi: 10.1109/IV.1997.626531

Gavish, N., Gutiérrez, T., Webel, S., Rodríguez, J., Peveri, M., Bockholt, U., & Tecchia, F. (2015). Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments, 23*(6), 778-798.

Haythornthwaite, C., Andrews, R., Fransman, J. & Meyers, E.M. (2016). Introduction to the sage handbook of e-learning research, Second edition. In C. Haythornthwaite, R. Andrews, J. Fransman, & E.M. Meyers (Eds.), The SAGE Handbook of E-learning Research, Second edition (pp. 3-20). Thousand Oaks, CA: SAGE.

Holton, E.F., III. (1996). The flawed four-level evaluation model. *Human Resources Development Quarterly, 7*(1), 5-21.

Issenberg, S.B., McGaghie, W.C., Petrusa, E.R., Gordon, D.L., & Scalese, R.J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher, 27*(1), 10–28.

Kalawsky, R.S. (1993). Critical aspects of visually coupled systems. In R.A. Earnshaw, M.A. Gigante, and H. Jones (Eds.), *Virtual Reality Systems* (pp. 203-212), San Diego, CA: Academic Press.

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

Kirkpatrick, J.D. & Kirkpatrick, W.K. (2016). *Four levels of training evaluation.* Alexandria, VA: ATD Press.

Lacerenza, C.N., Burke, C.S., Metcalf, D.S., Marlow, S.L., Read, L., Allen, C., & Mazzeo, M. (2015). Using augmented reality to train combat medics: An evaluation. *Proceedings from the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2015*. No. 15266.

Lathan, C. E., Tracey, M. R., Sebrechts, M. M., Clawson, D. M., & Higgins, G. A. (2002). Using virtual environments as training simulators: Measuring transfer. In. K. Hale & K.M. Stanney (Eds.), *Handbook of virtual environments: Design, implementation, and applications* (pp. 403-414). Mahwah, NJ: Lawrence Erlbaum Associates.

Lin, T. J., Duh, H.B.L., Li, N., Wang, H.Y., & Tsai, C.C. (2013). An investigation of learners' collaborative knowledge construction performances and behavior patterns in an augmented reality simulation system. *Computers and Education, 68*, 314–321

Lindner, M. (2007). What is microlearning? In M. Lindner & P.A. Bruck, (Eds.), Micromedia and corporate learning, Proceedings of the 3rd international microlearning 2007 conference (pp. 52-62). Innsbruck, Austria: Innsbruck University Press.

Loomis, J. M., Blaskovich, J. J., & Beall, A. C. (1999). Immesive Virtual Environment Training as a Basic Research Tool in Psychology. *Behavior Research Methods, Instruments, & Computers, 31*(4), 557–564.

MachCallum K. & Parsons, D. (2016). A thery-ology of mobile learning: Operationalizing learning theories with mobile activities. In L.W. Dyson, W. Ng & J. Fergusson (Eds.), *Mobile learning futures - Sustaining quality research and practice in mobile learning, Proceedings of the 15th World Conference on Mobile and Contextual Learning, mLearn 2016* (pp. 173-182). Sydney, Australia: University of Technology.

Mathieu, J.E., Tannenbaum, S.I., & Salas, E. (1992). Influences on individual and situational characteristic on measures of training effectiveness. *Academy of Management Journal, 35*, 828-847.

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education, 70*, 29-40.

Passmore, J. & Velez, M.J. (2012). SOAP-M: A training evaluation model for H.R. *Industrial & Commercial Training, 44*(6), 3215-325.

Russ-Eft, D., & Preskill, H. (2009). *Evaluation in organizations: A systematic approach to enhancing learning, performance, and change* (2nd ed.). Philadelphia: Basic Books.

Salas, E., Bowers, C. A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *The International Journal of Aviation Psychology, 8*(3), 197-208.

Schmidt-Daly, T.N., Riley, J.M., Hale, K.S., Yacht, D., & Hart, J. (2016). *Augmented REality Sandtable's (ARES) impact on learning* (Contractor Report No. ARL-CR-0803). Adelphi, MD: U.S. Army Research Laboratory.

Shams, L. & Seitz, A.R. (2008). Benefits of multisensory learning. *Trends in Cognitive Science, 12*(11), 411-417

Sharples, M. (2009). Methods for evaluating mobile learning. In G.N. Vavoula, N. Pachler, & A. Kukulska-Hulme (Eds.), *Researching mobile learning: Frameworks, tools and research designs* (pp. 17-39). Oxford: Peter Lang Publishing Group.

Sotomayor. T.M. & Alban, A. (2015). Augmenting training of the humeral head intraosseous (IO) procedure with a high-fidelity anatomical model. *Proceedings from the MODSIM World 2015 Conference.* No. 6.

Stone, R. T., Watts, K., & Zhong, P. (2011). Virtual Reality Integrated Welder Training. *Welding Journal, 90*(7), 136-S–141-S.

Stroud, K. J., Harm, D. L., & Klaus, D. M. (2005). Preflight virtual reality training as a countermeasure for space motion sickness and disorientation. *Aviation Space and Environmental Medicine, 76*(4), 352–356.

Tamkin, P., Yarnall, J., & Kerrin, M. (2002). Kirkpatrick and Beyond, A review of training evaluation. IES Research Network [Report 392]. Brighton, UK: The Institute for Employment Studies.

Tang, A., Owen, C., Biocca, F. & Mou, W. (2003). Comparative effectiveness of augmented reality in object assembly. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 5*(1), 73-80.

Underwood, J.S., Kruse, S., & Jakl, P. (2010). Moving to the next level: Designing embedded assessments into educational games. In P. Zemliansky & D. Wilcox (Eds.), *Design and Implementation of Educational Games: Theoretical and Practical Perspectives* (pp. 126-140). Hershey, PA: IGI Global. doi: 10.4018/978-1-61520-781-7

Webel, S., Bockholt, U., Engelke, T., Gavish, N., Olbrich, M., & Preusche, C. (2013). An augmented reality training platform for assembly and maintenance skills. *Robotics and Autonomous Systems, 61*(4), 398–403.

Webster, R. D. (2014). *Corrosion prevention and control training in an immersive virtual learning environment* (Technical Report No. NACE-2014-3726). San Antonio, TX: NACE International.