

Rethinking Effectiveness Evaluations: Measuring the Effectiveness of a Mobile Performance Support Application Using xAPI

Jennifer Murphy, Frank Hannigan, Tarah Daly

Quantum Improvements Consulting

Orlando, FL

jmurphy, fhannigan, tdaly@quantumimprovements.net

Chad Udell

Float Mobile

Morton, IL

chadu@gowithfloat.com

ABSTRACT

Evaluating the effectiveness of learning technology typically involves determining the extent to which a learner has gained and can apply knowledge, with an emphasis on long-term retention. Research designs in these studies typically involve one of the following methods: comparing a group's performance before and after an instance of training; comparing a group who receives training to a group who does not; or comparing groups who have received different formats or levels of the same training. While these designs may be relevant for evaluating a single instance of training, they do not effectively address how people learn in the real world; learning is a process that occurs over time with repeated exposure to material across a variety of experiences and platforms. That learning can occur anytime and anywhere is the foundational premise behind the Experience API (xAPI), a specification designed to enable tracking and managing human performance data across training platforms and instances. One example of how typical training effective research models fail involves evaluating the effectiveness of mobile learning. The benefits of mobile platforms include the ability to provide "just-in-time" training and performance support. The goal is not to support long-term retention of knowledge, but to provide just enough information at the point of need for the user to execute a task or solve a problem. This is a fundamentally different research question and requires evaluations designed to determine the extent to which the user can apply information in real time. In this paper, we will discuss how xAPI data from a mobile learning application can be used to determine its effectiveness. To illustrate the methodology, we will present data from a recently conducted effectiveness evaluation of a mobile application designed to support Warfighters' financial literacy.

ABOUT THE AUTHORS

Jennifer Murphy, PhD., is the founder and CEO of Quantum Improvements Consulting, LLC. She has 15 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. Her research has been featured in *The New York Times*, the *Pentagon Channel*, *Soldier Magazine*, and *Signal Magazine*.

Frank Hannigan is the Chief Operating Officer of Quantum Improvements Consulting, LLC. He has over 12 years of experience applying human factors and systems engineering principles and methods to the design of expert system displays and training curricula. Frank holds a M.S. in Human Factors & Systems from Embry-Riddle Aeronautical University and B.S. in Aviation Science from Saint Louis University. Frank has led projects on high-profile programs including the Navy's Zumwalt Class Destroyer (DDG-1000), the FAA's Next Gen Air Traffic Control System, has directly supported the DOE's Electric Grid Modernization efforts, and was tapped to execute a special assignment for the executive leadership at Sandia National Laboratories. Frank's research interests are rooted in human performance with respect to automation increases and technology enhancements, whether on the battlefield, in the cockpit, or in the classroom.

Tarah Daly is a Senior Human Factors Engineer at Quantum Improvements Consulting, LLC. She holds a BS in Psychology and an MS in Modeling and Simulation from the University of Central Florida. Tarah has worked on government-funded projects including assessing training of vigilance tasks in video game-based environments,

creating cognitive assessments to gauge skill decay, evaluating effectiveness of novel training technology to be utilized by the modern Warfighter, and researching psychophysiological paradigms for the investigation of a stress classifier utilizing mobile applications and technology. As an experienced human factors researcher, Tarah has designed and administered experimental protocol as well as conducted data analyses and has published work for DoD clients, academia, and industry.

Chad Udell is Managing Director at Float, LLC. Chad provides thought leadership to Float's programs and research and ensures the overall success of the organization and all client engagements. Chad guides business development, product design and development and creation of all of Float's deliverables. He has deep expertise in mobile design and development for learning and performance support, and has authored two influential books on the topic of mobile learning. His latest book, *Shock of the New: The Challenge and Promise of Emerging Learning Technologies*, provides a framework for evaluating technologies for use in organizational learning and development. Chad is a published industry expert and speaks regularly at national and international conferences on topics related to learning, mobility and design technology.

Rethinking Effectiveness Evaluations: Measuring the Effectiveness of a Mobile Performance Support Application Using xAPI

Jennifer Murphy, Frank Hannigan, Tarah Daly
Quantum Improvements Consulting
Orlando, FL
jmurphy, fhannigan, tdaly@quantumimprovements.net

Chad Udell
Float
Morton, IL
chadu@gowithfloat.com

INTRODUCTION

A perennial issue in the Defense training research community is the requirement to demonstrate the value of the training applications of technology in which we are so heavily invested. The usefulness of this technology is most typically discussed in terms of training effectiveness, which refers to the extent to which knowledge and skills are gained as the result of an intervention. Fundamentally, training effectiveness evaluations (TEEs) are designed to address the question of whether the trainee learned what they are supposed to learn. As such, an essential component of evaluation is to collect data directly related to learning objectives. Other training outcomes, such as trainee attitudes and skill transfer, are also often of interest. Researchers have been conducting TEEs for decades in military settings, and ample guidance exists for how to design experiments to address these research questions (e.g., Boldovici et al., 2002; Fletcher & Chatelier, 2000). Yet, regardless of the technology platform of interest, the results of TEEs tend to be mixed, and findings are rarely replicated. Authors have attempted to address this issue with the literature through meta-analyses (van Wijk et al., 2008; Arthur et al., 2003), review articles (Baldwin et al., 2009; Grossman & Salas, 2011), and conceptual frameworks (Kraiger et al., 1993), however, discrepancies remain.

Research into the effectiveness of mobile platforms for learning exemplifies the issues surrounding TEEs of technology-based training. Intuitively, mobile learning makes sense; mobile devices provide additional opportunities for learning outside the classroom, offer seamless access to information, and encourage personalized, collaborative, and informal learning (Baran, 2014). There is evidence that learners value, enjoy, and participate in mobile learning (Wong, 2017). While there is ample evidence that mobile technology can provide effective learning experiences (Wu, et al., 2012), some studies show no benefit to mobile learning above and beyond traditional methods (Marzouki et al, 2017; Lee et al., 2014). Further, when implemented inappropriately, mobile technology can prove detrimental to learning (Hwang & Wu, 2014). Taken together, these findings suggest that when done well, mobile learning is effective, but what makes it “good” is still a topic for research. However, methodological inconsistencies, lack of rigor, confusion of terminology, and other problems with the literature cloud the issue.

In addition to these issues, we propose that the types of experimental designs used in traditional TEEs inadequately reflect the way users interact with training technology applications. In this paper, we discuss a different way of conceptualizing the effectiveness of technology applications in the context of evaluating a mobile learning application. Our approach is different from traditional approaches to TEE in two ways. First, we consider a key affordance of mobile technology: the ability to have access anytime, anyplace to relevant information. Second, instead of conducting an evaluation on recall after a single instance of training, we consider the way users typically interact with mobile applications, which is as a tool used repeatedly over time.

In this paper, we will discuss how technology-based training, and mobile learning in particular, have historically been evaluated. Then, we will describe the approach to evaluating a mobile learning application our team has been developing to support service members’ financial literacy. Importantly, we will discuss how implementing the Experience API (xAPI) specification within a mobile application enables development of objective, unobtrusive measures of usability and effectiveness. Finally, we will discuss findings from the first of several experiments to be conducted with this application to demonstrate the validity of our evaluation approach and describe the path forward.

LIMITATIONS OF CURRENT APPROACHES TO TRAINING EFFECTIVENESS

While the pace of advances in training technology has accelerated, the experimental designs and approaches used to evaluate these applications have not changed much over the past 50 years. Most training effectiveness researchers are familiar with Kirkpatrick's "Four Levels" of evaluation (Kirkpatrick, 1959; 1966), including the revised "New World Model" (Kirkpatrick & Kirkpatrick, 2016). This approach addresses training effectiveness based on four criteria, or levels (Reactions, Learning, Behavior, Results). While this remains the most popular model, it is by no means the only one currently in use. Another well-known model in the Defense industry is the Integrated Model of Training Evaluation and Effectiveness (IMTEE), proposed by Alvarez et al. (2004). This model is unique in its consideration of both training evaluation – the method by which training provides the required learning outcomes – and training effectiveness – a theoretical approach to the underlying psychology of why these learning outcomes were or were not achieved. Passmore & Velez (2014) review 11 models of training effectiveness and their specified criteria. Many of the models they reference were developed by human resources professionals to evaluate coaching or leadership programs, courses, or classes which do not rely upon technology. Describing these models in depth is outside the scope of this paper. However, most of them share characteristics that prove challenging when attempting to put the models into practice.

Current models of training effectiveness are prescriptive regarding the sorts of outcomes a practitioner should assess, but do not provide specific guidance on when or how to measure these outcomes. Most consider improvements on the organizational level (e.g., productivity, profitability) to be the most important outcome of training. While improving the effectiveness of an organization is typically the ultimate goal of training programs, evaluating this facet of training effectiveness often proves elusive. Consequently, many TEEs focus on constructs such as satisfaction with the learning experience or knowledge gained. When researchers attempt to implement these models in applied settings, where participant time is hard to come by, data collection is messy, and sample sizes are small, often beautifully constructed research designs devolve into simple pre- and post-test comparisons on an unvalidated multiple choice test.

Drawing conclusions about the effectiveness of mobile learning based on the literature is challenging, largely because "mobile learning" has been operationally defined as a function of the mobility of the learner versus the platform itself (Valvoulas & Sharples, 2008). Literature reviews include a wide variety of technologies, from laptops to cameras to podcasts (Wu et al., 2012). Further, mobile platforms support a broad range of instructional strategies across many domains for audiences including K-12, higher education, corporate training, and defense training and education. A variety of measures of effectiveness are employed, including learning achievements, perceptions of usefulness, motivation, cognitive load, and attitudes (Wong, 2018). In most cases, these outcomes are evaluated through a survey methodology immediately following training. In cases where experiments are conducted, the control condition is usually a more traditional instructional method (e.g., textbook, instructor led course). While the gold standard of TEEs involves some sort of between group evaluation, often the differences between traditional and mobile learning methods are such that this comparison is somewhat of a straw man. This particular issue is not limited to mobile learning; as we evaluate emerging technologies for learning such as AR and VR, we need to consider that if these technologies are appropriately implemented, learning experiences will be sufficiently different from traditional approaches as to render "apples to apples" comparisons inappropriate.

Another problem with TEEs as they are currently conducted has to do with how we think about what we call "training technology." As a field, our research has focused on technologies such as computer-based learning (eLearning), simulations, and other platforms in which applications are designed with the goal of increasing domain knowledge and providing opportunities for practicing skills. Once a level of proficiency is reached, the trainee purportedly has learned what he or she was supposed to learn and using the application on a regular basis is no longer necessary. In these cases, it may be appropriate to evaluate a singular instance of training, as is typical of most TEE research designs. However, newer technologies of interest, such as extended reality, intelligent assistant devices, and smartphones, provide opportunities to enhance performance outside of just learning. While warfighters can certainly learn on these platforms, they are not "training technologies" per se; the goal is to provide real time performance support. If an augmented reality application is designed for training, eventually the warfighter operates without the headset. On the other hand, if the application is designed for performance support, the warfighter will keep the headset on because it is providing information in real time he or she no longer is required to learn. Our field has seen an evolution toward this way of thinking with the advent of the smartphone. As researchers saw the potential for mobile devices to provide anytime, anyplace access to information, the phrase "just-in-time training" emerged. What has become apparent is

that providing “just-in-time” content is not the same as training; we no longer have to learn directions in a new city, because there’s an app for that.

If we consider how warfighters interact with technology with an increasing focus on performance support, evaluating the effectiveness of these applications requires a shift in what we measure and how. Instead of asking “Does a group trained in a simulator produce higher qualification scores than a group that is not?” research questions should reflect the extent to which operational performance is facilitated with the technology (e.g., “Does a group with an augmented reality wayfinding application complete a land navigation exercise faster than one that does not?”). Questions about usability become critically important, as the extent to which the user can easily and quickly interact with the application’s interfaces determines access to information. To fully capture the extent to which an application supports performance, measures of performance should be collected repeatedly and over time, versus traditional designs in which training is considered a singular event.

USE CASE: EVALUATING A FINANCIAL LITERACY MOBILE APPLICATION

To facilitate discussion of how the effectiveness of technology applications for learning versus performance support differ, we will provide an example in the context of a mobile application. Under an effort funded by the Office of Financial Readiness and the ADL Initiative, Sen\$e is a smartphone (Android and iOS) application designed to increase service members’ financial literacy. The application is currently undergoing beta testing with the aim of being released, free of charge, on the Apple and Android application marketplaces by the end of 2019.

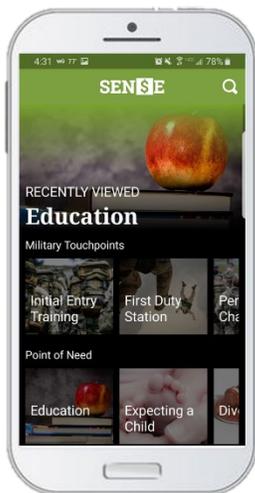


Figure 1: Sen\$e Launch Screen

The primary impetus for the application was 10 U.S. Code § 992, which requires financial literacy training to be provided to service members at distinct times throughout their military career, or “touchpoints.” Touchpoints include events such as First Duty Station, Promotion, and Permanent Change of Station. To address the financial issues specific to each touchpoint, relevant microlearning content is presented in a section devoted to each one (See Figure 1). In addition, content is organized into “points of need,” which support finding the answers to specific financial questions that a warfighter may have at any point in their career. These points of need include marriage, homebuying, expecting a child, and other infrequent events. In addition, the application features a glossary, search functionality, links to external sources of information, and mini-games. Microlearning content in the application was developed from government sources and reviewed by a panel of financial experts across the Services, Department of the Treasury, Office of Personnel Management, and other Federal government departments.

To evaluate the effectiveness of Sen\$e, our team conducted an evaluation with a prototype version of the application focusing on usability and the extent to which participants could find the answers to financial questions in real time. In addition, we evaluated learning using a knowledge test. While this method did evaluate the application as a performance support tool, addressing the issue of evaluating performance across multiple exposures to the application was not possible without implementing the Experience API (xAPI) into the application. These findings will be described below, and how xAPI was implemented to address future research will be discussed.

METHODOLOGY

Participants

Twenty-eight participants from across four branches of the military (U.S. Navy, U.S. Air Force, U.S. Marine Corps, and U.S. Army) stationed at Joint Base Pearl Harbor-Hickam, Marine Corps Base Hawaii, and Schofield Barracks, respectively ($M_{\text{age}} = 32.18$ years; $SD = 7.63$; men = 18, women = 10) participated in this study. All participants were serving in the military at the time of participation ($M_{\text{service}} = 10.23$ years; $SD = 6.77$), and each had a different career field within their respective service. The qualifiers for participation were to be at least 18 years of age, current or former service member, and had no previous experience with the Sen\$e application. Participants were instructed that their participation was voluntary, and that the evaluation could end at any time.

Procedure

Participants met with a study facilitator and note-taker at their respective military installation. Participants first read and signed the informed consent form and were introduced to the goals of the study. Next, participants completed a demographics questionnaire. Participants were then invited to take five minutes to freely explore the Sen\$e application. The study facilitator set a timer for the five-minute duration and participants were handed the phone of their choice. Participants used the type of phone they indicated they were most comfortable using for the duration of the study (iPhone 7 Plus, iPhone 6s, Samsung Galaxy S7, or Google Pixel). A beta version of the application was presented on one of the four phones.

After the five-minute exploration, participants reviewed a total of five content areas, one area at a time, and completed a paper-based multiple-choice test for each content area. Participants were able to use the application while answering the questions, if needed. At the end of the evaluation, participants were asked to respond verbally to several debrief questions regarding their experience with the application. Participants were then thanked for their time and dismissed.

Test Scenarios

Test scenarios were identified by the project team based on available functionality of the application and reflected tasks which included a variety of in-application interactions and financial content. These scenarios required participants to navigate to key categories on the main dashboard like “Point of Need” and “Touchpoints.” The test scenarios were as follows:

Scenario 1: “Navigate to the ‘Divorce’ section. Here, you will review content from the ‘During the Divorce’ area.”

Scenario 2: “Navigate to the ‘Birth of a Child’ section. Here, you will review content from the ‘Post-Delivery’ area.”

Scenario 3: “Navigate to the ‘Disability’ section. Here, you will review content from the ‘Immediately After an Injury’ area.”

Scenario 4: “Navigate to the ‘Basic Finance’ section. Here, you will review content from the ‘Cash Flow’ area.”

Scenario 5: “Navigate to the ‘Vehicle Purchasing’ section. Here, you will review content from the ‘Cost of Owning a Vehicle’ area.”

Each scenario required participants to access and review content in a different content category from the dashboard. After reviewing each content area, participants answered three to four questions, and participants were allowed to use their devices while taking the challenges, if needed. In total, seventeen questions were answered. All questions were scenario-based and were cross-validated for applicability and difficulty by at least two other members of the project team.

RESULTS

Participant responses were recorded and graded across each of the 17 questions within each of the five content sections (Divorce, Birth of a Child, Disability, Basic Finance, and Vehicle Purchasing). Data are presented across all participants, by military branch, and by age range. Age ranges were selected to capture equal numbers in each range.

Overall, participants scored an average of 82.57% ($SD = 17.05$) correct, with 71.60% ($SD = 33.20$) correct responses in the “Divorce” content, 93.83% ($SD = 2.14$) correct responses in the “Birth of a Child” content, 81.48% ($SD = 13.52$) correct responses in the “Disability” content, 100% ($SD = 0$) correct responses in the “Basic Finance” content, and 70.37% ($SD = 29.78$) correct responses in the “Vehicle Purchasing” content. See Figure 2 for results.

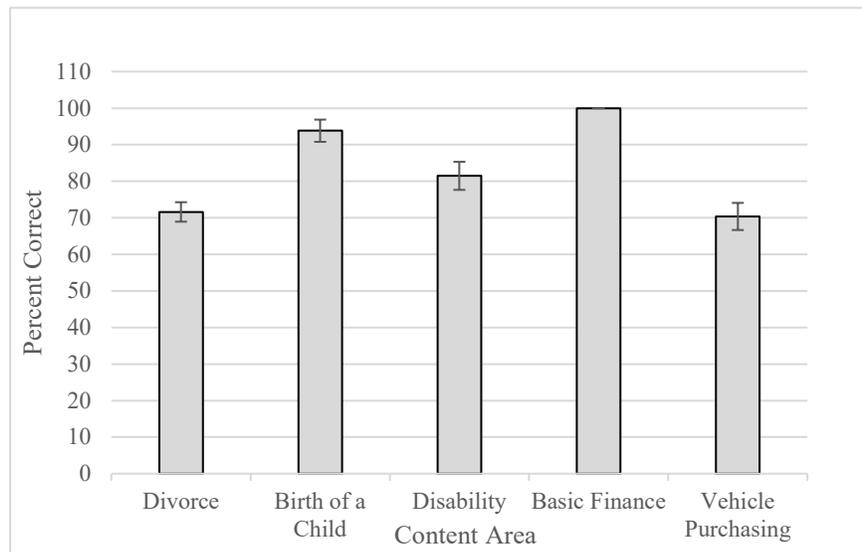


Figure 2. Knowledge test performance by content area

While these findings suggest participants can find relevant information in the application when it's needed, this approach does not speak to how end users would ultimately use Sen\$e outside of an experimental setting. To address the goal of evaluating the application as a performance support tool, collecting usage data is required. While our ability to collect data over time is dependent upon the release of the application, here we describe the development to enable it.

IMPLEMENTATION OF xAPI

In the past 10 years, the Experience API (xAPI) has evolved from a concept for the “next generation of SCORM” to a specification used in the Defense and corporate learning spaces to track performance data. The primary benefit of using the xAPI specification is the ability it affords to store human performance data from multiple sources in a single, intuitive format. Because of its flexibility, xAPI enables the capture of a wide variety of learning experiences, both inside and outside the classroom. In terms of TEEs, there are a number of implications. xAPI supports the development of robust, persistent learner models in training systems. As a result, it is possible to track performance across multiple training. Importantly, all types of experiences can be represented in xAPI format, including events that occur completely outside of a training environment. Whereas evaluations of transfer performance were previously limited in terms of reliable access to operational performance data, xAPI enables objective assessment of skill transfer to a higher fidelity or live scenario.

While xAPI has been implemented in a variety of learning applications, often data are collected but not used by stakeholders to make decisions about the effectiveness of their investments. Typically, data are used to track completion of training requirements, accuracy of test responses, and other broad level analytics. Our approach to implementing xAPI is different in that we aim to address whether the Sen\$e application supports end users when and where they need support. To that end, we developed a data collection capability within the application to support continuous usability and effectiveness evaluation.

The first step in this process was to establish research questions. Through discussions with stakeholders, a number of outcomes of interest were identified. These include rates of usage by targeted demographic groups, validation of the content structure, relevance of content to specific touchpoints, and other issues. Internal to our team, we identified usability questions that could be answered through patterns of usage. For example, if features or content areas are rarely leveraged, ease of access may be an issue. In particular, we are interested in the learning path a user takes

through the content. By tracking performance data, we can address questions such as whether users tend to complete an entire content area in one sitting, how long it takes to complete a section, and the order in which content is accessed.

Once research questions were established, our team developed measures to address each one. These measures consist of a combination of demographic and usage data within the application. Demographic data is collected through a survey administered during profile set up, and includes questions about users' age, location, goals for using the application, topics of interest, branch, rank, and time in service. Usage data collected within the application required the identification of specific activities within the user flow to trigger the push of an xAPI statement to the learner record store. From these statements, our team can extrapolate measures of behavior over time. An example of an xAPI statement from Sen\$e is below:

```

{
  "actor": {
    "objectType": "Agent"
    "account": {
      "homePage": "https://sense.usalearning.net/"
      "name": "532d19c8-dfa9-4590-bf0e-4f8e85112d67"
    }
  }
  "verb": {
    "id": "http://adlnet.gov/expapi/verbs/launched"
    "display": {
      "en-US": "launched"
    }
  }
  "object": {
    "objectType": "Activity"
    "id": "tag:adlnet.gov,2013:expapi:0.9:activities:ofr-frpal://activity/3f27d2c4-49a7-44f6-a63b-e7e2562e3727"
    "definition": {
      "type": "http://adlnet.gov/expapi/activities/course"
      "name": {
        "en-US": "DTI Ratio Calculators"
      }
    }
  }
  "context": { ... }
  "timestamp": "2019-06-20T14:23:02.094Z"
  "id": "6c620e88-6e06-4115-8be1-26a270c29044"
  "version": "1.0.1"
  "stored": "2019-06-21T12:50:22.528Z"
  "authority": {
    "objectType": "Agent"
    "account": {
      "homePage": "https://sense-production.lrs.io/keys/prod"
      "name": "prod"
    }
  }
}

```

Figure 3: Sample xAPI statement from Sen\$e

The statement structure captures the behavior of a user launching one of the calculators in Sen\$e. The user's identifying information, activity, timestamp, and contextual information such as course title are recorded. Using xAPI to track users' performance enables us to address a broad range of questions. These include:

- What are the most frequently searched for topic areas?
- Do younger users play the games within the application more than older ones?
- Do users access the content through the touchpoints or through the search functionality?
- How long is the average usage session?
- Do users access optional content?
- Do users complete post-content assessments?

Usage patterns will enable our team to identify and remediate usability issues, evaluate the usefulness of content to specific demographics, and evaluate the effectiveness of various functions and features in the application. Importantly,

these data can speak to broader research questions about mobile learning, performance support, and self-regulated learning.

Our broader research goal is to contribute to the growing body of research aimed at evaluating “microlearning.” To date, microlearning is a poorly defined construct that generally refers to providing “bite-sized” chunks of information versus longer form didactic content. In theory, microlearning and mobile learning go hand in hand, as the form factor of mobile devices supports shorter, targeted learning sessions with multimedia content. However, there is little industry consensus on what qualifies as “microlearning.” How long is an average “microlearning” session? While practitioners agree the answer is “just long enough,” there are no data that speak to how long a user tends to spend in a self-directed learning context on a mobile device. Usage data gathered through xAPI statements can help address these questions.

Obviously, best practices will depend upon a variety of factors including content type, motivation, and access, but determining what works will be challenging if our effectiveness research continues to revolve around traditional TEE research designs such as between group comparisons or pre-post within subjects designs using knowledge tests as performance criteria. By evaluating performance trends over time, learning paths, and group differences, as a community we can forge new ground in evaluating mobile learning, and technology-based training more broadly.

REFERENCES

- Alvarez, K., Salas, E., & Garofano, C. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review*, 3(4), 385–416. doi:10.1177/1534484304270820
- Arthur, W., Bennett, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology*, 88(2), 234–245. doi:10.1037/0021-9010.88.2.23
- Baran, E. (2014). A review of research on mobile learning in teacher education. *Educational Technology & Society*, 17 (4), 17–32.
- Boldovici, J. A., Bessemer, D. W., & Bolton, A. E. (2002). *The elements of training evaluation*. Ft Belvoir, VA:U.S. Army Research Institute. Retrieved from <http://www.dtic.mil/get-tr-doc/pdf?AD=ADA402961>
- Fletcher, J. D., & Chatelier, P. R. (2000). An overview of military training, (Report 2514). Alexandria, VA:Institute for Defense Analysis. Retrieved from <http://www.dtic.mil/get-tr-doc/pdf%3FAD%3DADA408439&hl=en&sa=T&oi=gsb-gga&ct=res&cd=0&ei=y64dWqTuFdSLmAG-gLSwAw&scisig=AAGBfm23bJ7tkcXzsaU71p8k0OAD1oJbma>
- Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development*, 15(2), 103–120. doi:10.1111/j.1468-2419.2011.00373.x
- Hwang, G. J., & Tsai, C. C. (2011). Research trends in mobile and ubiquitous learning: a review of publications in selected journal from 2001 to 2010. *British Journal of Education Technology*, 42(4), E65–E70.
- Kirkpatrick, D. L. (1967). Evaluation of training. In R. L. Craig and L. R. Bittel (Eds.) *Training and development handbook*. New York, NY: McGraw-Hill Publishers.
- Kirkpatrick, J., & Kirkpatrick, W. (2015). *An introduction to the new world Kirkpatrick model*. Retrieved from <http://www.kirkpatrickpartners.com/Portals/0/Resources/White%20Papers/Introduction%20to%20the%20Kirkpatrick%20New%20World%20Model.pdf>
- Kraiger, K., Ford, K., & Salas, E. (1993). Application of cognitive, skill-based and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78(2). doi:0031-9010/93/
- Lee, S., Barker, T., & Suresh Kumar, V. (2016). Effectiveness of a learner-directed model for e-learning. *Educational Technology & Society*, 19 (3), 221–233
- Passmore, J., & Velez, M. J. (2014). Training evaluation. In K. Kraiger, J. Passmore, N. Santos, S. Malvezzi (Eds.), *The Wiley Blackwell handbook of the psychology of training, development, and performance improvement* (pp. 136–153). Hoboken, NJ: John Wiley and Sons
- Marzouki, O., Idrissi, M., & Bennani, S. (2017). Effects of social constructivist mobile learning environments on knowledge acquisition: A meta-analysis. *Interactive Mobile Technologies* 11(1) 18-39.
- van Wijk, R., Jansen, J. J. P., & Lyles, M. A. (2008). Inter- and intra-organizational knowledge transfer: A meta-analytic review and assessment of its antecedents and consequences. *Journal of Management Studies*, 45(4), 830–853. doi:10.1111/j.1467-6486.2008.00771.x
- Vavoula, G. N., & Sharples, M. (2008). Challenges in evaluating mobile learning. Paper presented at the annual meeting of the mLearn Conference, Wolverhampton, United Kingdom.

- Wong, B. (April, 2018). Success in mobile and ubiquitous learning: Indicators of effectiveness. *Broad Research in Artificial Intelligence and Neuroscience* 9, Special Issue on Educational Psychology.
- Wu, W.-H., Wu, Y.-C. J., Chen, C.-Y., Kao, H.-Y., Lin, C.-H., & Huang, S.-H. (2012). Review of trends from mobile learning studies: A meta-analysis. *Computers & Education*, 59(2), 817–827.