How Can We Measure Learning? Let's Count the Ways!

Jeffrey M. Beaubien, Ph.D.[1], E. Webb Stacy, Ph.D.[1], Sterling L. Wiggins, M.A.[2], Lisa C. Lucia, Ph.D.[1]
Aptima, Inc.
Woburn, MA[1], Fairborn, OH[2]
jbeaubien@aptima.com, wstacy@aptima.com, swiggins@aptima.com, llucia@aptima.com

ABSTRACT

Although similar, learning and performance are not synonymous. Specifically, performance is defined as the quality, rate, or accuracy of a specific behavioral response at a specific point in time. By comparison, learning is defined as a relatively permanent change in knowledge, skills, or understanding across tasks, across time, and/or across environments (Christina & Bjork, 1991). While this distinction is applicable to all training methods, it is particularly relevant in the context of simulation-based training (SBT), because instructors need to ensure that learners have mastered the critical work-related skills, rather than simply having learned "how to game the simulator." In a recent I/ITSEC paper (Beaubien, Stacy, Wiggins, Keeney, Walwanis et al., 2015), measures of learning (MOL) were conceptually and empirically differentiated from measures of performance (MOP) during simulation-based carrier landing practice. The MOLs and MOPs provided vastly different interpretations of the data. In particular, it was shown that a singular reliance on MOPs would have provided the Navy with incorrect guidance regarding the design of future flight simulators for training carrier landing skills. This follow-on paper explores nine theoretically-derived methods for measuring learning. Several are rooted in the dual-process theory of decision-making (Evans & Stanovich, 2013, Kahneman, 2011), which postulates two brain processes – an analytical one which is characteristic of novices, and an intuitive one which is characteristic of experts – that operate in parallel. The remaining methods are rooted in a generalized model of skill acquisition (Dreyfus & Dreyfus, 1980), which postulates that there are predictable patterns of cognitive and behavioral development over time as one becomes more expert-like. Each assessment method is described using lay terminology so that non-scientists can integrate them into their own training efforts, SBT or otherwise. Specific examples from the published literature are also used to illustrate key points.

ABOUT THE AUTHORS

Jeffrey M. Beaubien, Ph.D. is a Principal Scientist at Aptima, Inc., where he leads projects on training and human performance assessment in high-risk environments. His research interests include team dynamics, adaptability, and decision-making. Dr. Beaubien received a Ph.D. in Industrial and Organizational Psychology from George Mason University, a M.A. in Industrial and Organizational Psychology from the University of New Haven, and a B.A. in Psychology from the University of Rhode Island.

E. Webb Stacy, Ph.D. is a Corporate Fellow at Aptima, Inc., where he is responsible for enhancing Aptima's science and technology portfolio. Dr. Stacy has an interest in using modern Cognitive Science to improve experiential training. His recent work includes investigating the relationship of simulator fidelity to training effectiveness, and developing an approach for optimizing the training value of experiential scenarios. Dr. Stacy holds a Ph.D. in Cognitive Science from SUNY/Buffalo, and a B.A. in Psychology from the University of Michigan. He is a past Program Chair for the Society for Behavior Representation in Modeling and Simulation (BRIMS).

Sterling L. Wiggins, M.A. is a Principal Scientist at Aptima, Inc. He leads several projects that develop technology and training solutions for operators in high-risk, safety-critical environments. His research interests include Live, Virtual, and Constructive (LVC) training, human-automation interaction, and adaptive aiding. He holds an M.A. in Education with a focus on learning, design, and technology from Stanford University, and a B.S. in Psychology from Ohio State University.

Lisa C. Lucia, Ph.D. is a Scientist at Aptima, Inc. She leads projects that involve the development of neuroscience-based training tools, healthcare applications, cognitively-guided decision support tools, and cortical motor control

systems. Her research interests span from the brain-bases of human perception and memory to visuospatial abilities and skill learning through training. She holds a Ph.D. in Cognitive Neuroscience from Tufts University, and a B.S. in Biological Psychology from Bates College.

How Can We Measure Learning? Let's Count the Ways!

Jeffrey M. Beaubien, Ph.D.[1], E. Webb Stacy, Ph.D.[1], Sterling L. Wiggins, M.A.[2], Lisa C. Lucia, Ph.D.[1]
Aptima, Inc.
Woburn, MA[1], Fairborn, OH[2]
jbeaubien@aptima.com, wstacy@aptima.com, swiggins@aptima.com, llucia@aptima.com

INTRODUCTION

Although similar, learning and performance are not synonymous. Specifically, performance is defined as the quality, rate, or accuracy of a specific behavioral response at a specific point in time. By comparison, learning is defined as a relatively permanent change in knowledge, skills, or understanding across tasks, across time, and/or across environments (Christina & Bjork, 1991). Since they are not synonymous, it is possible to observe changes in performance without any corresponding changes in learning. For example, when cramming for an exam, learners often rely on superficial learning strategies such as reviewing class notes and re-reading passages of highlighted text. As a result, they experience a momentary reaction potential (Hull, 1943) which provides the illusion of learning. However, the absence of learning is clearly revealed after a period of disuse, for example when re-administering the same exam two weeks later. Not surprisingly, performance tends to drop substantially on the second exam, thereby demonstrating that learning did not occur. Conversely, it is possible to observe gains in learning without corresponding changes in performance. For example, when practicing a task to the point of automaticity, the learner's performance will plateau upon reaching the criterion of mastery. This provides the illusion that no further learning gains are occurring during the subsequent post-plateau practice trials. However, the true extent of learning is again easily demonstrated after a period of disuse, for example when comparing the relative skill decay and reacquisition rates of individuals who engaged in overlearning versus matched controls who did not. As a general rule, higher levels of overlearning tend to produce less decay and faster reacquisition, thereby demonstrating that learning continued even after performance had plateaued.

The critical distinction between learning and performance was recently demonstrated in a study on the effects of simulator fidelity cues on the acquisition of carrier landing skills (Beaubien et al., 2015). This study included a sample of fifteen Navy F/A-18 pilots (8 novices, 7 experts), each of whom flew 24 landing passes in a high-fidelity simulator over two consecutive days. Measures of Performance (MOPs) were calculated for each pass, and were operationalized as deviations (measured in degrees) from the ideal angle of attack, glide slope, and center line during the last 18-23 seconds of the final approach. The measures were then aggregated across all 24 landing passes to provide a single, average performance score for each participant. By contrast, Measures of Learning (MOLs) were operationalized as changes in performance across four "blocks" of passes during the two-day exercise. The two sets of analyses – learning vs. performance – provide very different interpretations of the data. Specifically, the MOLs showed that the enhanced fidelity cues initially degraded the novices' performance. However, their performance subsequently improved and eventually became statistically indistinguishable from the experts by the end of training. By comparison, the MOPs showed that the enhanced fidelity cues resulted in lower average performance. If the research team had only calculated the MOPs, they would have provided the Navy with incorrect guidance regarding the design of next-generation simulators for training carrier landing skills.

This follow-on paper explores nine theoretically-derived methods for measuring learning. Several are rooted in the dual-process theory of decision-making (Stanovich & West, 2013; Kahneman, 2011), which postulates that there are two brain processes – one rapid and intuitive, the other slow and analytical – that operate in parallel. The remaining methods are rooted in a generalized model of skill acquisition (Dreyfus & Dreyfus, 1980), which postulates that there are predictable patterns of cognitive and behavioral development with increasing expertise. These two theories are entirely consistent with one another; they merely differ in their primary focus area. Specifically, dual-process theory compares and contrasts the extreme ends of the skill acquisition continuum (novices vs. experts), while the generalized skill acquisition model focuses on qualitative changes across the various stages of mastery. While many of the assessment methods will be familiar to cognitive, human factors, and sports psychologists, they will likely be unfamiliar to training professionals who do not have formal training in psychology. As a result, the methods are described using lay terminology and presented using real-world examples, with the ultimate goal of enhancing their use throughout the modeling and simulation community-at-large. Specific examples from the published literature are also used to help illustrate key points.

THEORETICAL BACKGROUND

As noted previously, four of the assessment methods are rooted in the dual-process theory of decision-making, which postulates that there are two brain systems operating in parallel. Shorthand labels for these brain processes are "System 1" and "System 2," even though the labels technically describe different brain processes rather than brain systems (Stanovich & West, 2013). One of these processes, "System 1," operates at an unconscious level. This type of decision-making is extremely fast, makes minimal demands on working memory, and operates by associatively comparing the current situation to one's corpus of accumulated prior experiences stored in long-term memory. All humans engage in a considerable amount of System 1 processing in their day-to-day lives; in addition, experts often have well-developed System 1 skills in their specific domain of expertise. For example, a chess master can quickly look at a populated chess board, determine each player's strategy, and project the next two or three best moves – all within a matter of seconds. By comparison, "System 2" operates at the conscious level. It is much slower, places heavy demands on cognitive resources such as working memory, and makes decisions based on calculations and deliberation. It is akin to the slow and deliberate decision making approach used by novices, as well as the slow and deliberate approach used by experts when facing novel problems or situations for which their expertise has not prepared them (Kahneman & Klein, 2009). For example, a novice chess player must take time to carefully study the two players' positions, deliberately reflect on their relevant strategies, and consciously recall the relevant decision rules. Even with all this effort, novices can usually project only a single move forward.

The remaining five assessment methods are rooted in a generalized model of skill acquisition (Dreyfus & Dreyfus, 1980), which postulates that learners progress through several discrete stages of development: novice, competent, proficient, and expert[1]. Having no direct hands-on experience, *novices* understand the task or domain based entirely on what they have read in a book, heard in a lecture, or learned through observation. As a result, their knowledge is abstract and poorly-organized. During practice, their performance is slow and follows the textbook instructions; they must also consciously monitor their behavior to avoid making common errors, such as accidentally skipping a step. At the *competent* stage, learners have amassed a small corpus of prior experiences upon which to draw. While they can identify clear-cut cues that are consistently correlated with particular outcomes, their identification of subtle (and/or mixed) cues requires assistance from the instructor. Competent learners also rely on instructor-provided guidelines or heuristics to support their decision-making. However, since their behavior is not goal-directed, they tend to treat all guidelines as being equally important, regardless of the unique situational characteristics that might favor one or more guidelines over the others. With an even greater repertoire of experiences, learners eventually reach the *proficient* stage. Here, their behavior becomes goal-directed and contextualized. Using the goal or mission as their organizing framework, certain situational cues now take on greater meaning than others; similarly, they can now prioritize decision guidelines rather than treating them all as being equally important. However, they must still expend cognitive resources to consciously monitor their performance in real-time. They also have difficulty engaging in metacognitive activities such as dynamic re-planning. Finally, at the *expert* stage, learners have amassed an extensive repertoire of prior situations upon which to draw. Each situation now elicits a rapid, intuitive, and recognition-based decision. Experts' task-related performance tends to be smooth and efficient, often achieving economy of motion by integrating two or more discrete steps into a single, fluid one. Finally, because they no longer need to consciously monitor their performance in real-time, experts have reserve mental capacity which helps them to critically reflect on their own performance, dynamically re-plan, and change course on-the-fly (Dreyfus & Dreyfus, 1980).

In the following sections, we describe nine theoretically-derived methods for assessing learning. For each one, we describe its theoretical underpinnings, highlight some use cases in which it has been applied, and provide brief commentary to help training professionals incorporate the method into their own work. In practice, these nine learning assessment methods are employed in one of two general ways. The first is to measure the learner's performance immediately prior to and then again immediately following a block of instruction. The difference between the pre-test and post-test assessments is used to infer learning. The second is to measure the learner's performance on every practice trial, and then compute skill acquisition curves over time. Systematic changes in skill

[1] Since so few learners actually make it to the final stage of mastery, we have purposely excluded it from this discussion.

acquisition, decay, and/or reacquisition are used to infer learning. Both approaches have a long history of use in the psychological literature (Newell & Rosenbloom, 1981) and are equally valid. The decision to use one approach versus the other is often based on several factors, such as cost, practicality, and the specific question that one seeks to answer. For example, if one is interested in comparing two different instructional methods, the pre-test/post-test paradigm might be appropriate. However, if one is interested in determining optimal re-training intervals based on the rate of skill decay, then the learning curve paradigm might be appropriate.

MEASUREMENT METHODS INSPIRED BY THE DUAL-PROCESS THEORY OF DECISION-MAKING

In this section, we review four measurement methods that capitalize on well-known differences between "System 1" versus "System 2" decision processes. For example, "System 1" decision processes are much more automatic in their execution and require fewer cognitive resources, such as working memory.

Method #1: Decision Speed and Accuracy

Decision speed is defined by reaction time (RT) and is measured in milliseconds, with smaller numbers indicating better performance (because time is the inverse of speed). Decision accuracy is defined as the absolute correctness of a forced-choice decision in response to a pre-defined stimulus (Luce, 1986). In practice, this assessment method is typically employed using a pre-test/post-test assessment paradigm. For example, prior to training, participants may be presented with a series of still images or brief videos. Each stimulus appears on the screen for only a few seconds. The screen is then immediately masked, and the participant must make a rapid decision – using a mouse click, a keyboard press, or a trigger pull – about the stimulus that was just presented. The assessment process is then repeated again after the training is complete. For each learner, the average RT (in milliseconds) and decision accuracy (in percentages) are calculated to generate a baseline (pre-training) as well as a learning (post-training) score. The amount of learning is calculated as the difference between these two scores, and is typically analyzed using conventional statistical methods, such as a dependent t-test or Repeated Measures Analysis of Variance (RMANOVA). The amount of learning may be expressed in raw units, such as milliseconds or percent accuracy. Alternatively, it may be expressed in terms of effect size (Cohen, 1988), which divides the pre-test/post-test difference score by the pooled standard deviation, to facilitate comparison across studies or samples.

Generating the specific assessment stimuli can often be accomplished by connecting a commercial-off-the-shelf (COTS) video capture tool to the simulator, and then recording either still images or brief video clips from pre-scripted simulator runs. For example, in a submarine navigation exercise, a series of still images (as viewed through the periscope) might depict commercial vessels at various distances and angles from the submarine. For each image, the learner must quickly decide whether or not that vessel presents a potential collision threat, given the submarine's current heading and speed. Presenting and scoring the stimuli can also be relatively straightforward. Several PC-based software tools, such as *OpenSesame* (Mathôt, Schreij, & Theeuwes, 2012) and *PsychoPy* (Peirce, 2009), are freely available. These tools allow the trainer to set up the decision task, randomly order the images or videos, administer the decision task under controlled conditions, collect the RT and accuracy data, and export the data for subsequent statistical analysis. Measures of RT and accuracy have a long history of use in psychological research. For example, in a study of spatial skill training and transfer (Wright, Thompson, Ganis, Newcombe, & Kosslyn, 2008), two participant groups were initially tested on tasks of mental rotation, mental paper folding, and verbal analogies. Each group then participated in 21 days of practice for either mental rotation or mental paper folding. At the conclusion of training, the participants were again tested on all three tasks. Learning was defined as the change in average RT and accuracy over time, and was calculated separately for each task. As expected, the group trained in mental rotation showed a reduction in RT and percent error on the mental rotation task, while the group trained in paper folding showed a reduction in RT and percent error on the mental paper folding task. Neither group showed a statistical change in learning for their other tasks.

Method #2: Spare Working Memory Capacity

As decision-making moves from System 2 to System 1, working memory should also play a much smaller role in the decision-making process. The usual explanation is that learning involves "chunking" the knowledge or skill (Miller, 1956; Servan-Schreiber & Anderson, 1990; Gobet, Lane, Croker, Cheng, Jones, Oliver, & Pine, 2001). The amount of "spare" working memory capacity – and by extension, the extent to which the learner is relying on expert-like intuitive decision-making processes – can be quantified using the dual-task paradigm, of which the Peripheral Detection Task (PDT; Martens & van Winsum, 2000) is one example. In the PDT method, the learner is required to perform two tasks simultaneously. For example, the primary task might be to land a fighter jet on an aircraft carrier. The secondary might be to continuously monitor a blinking light that appears in the periphery of the user's visual field, and to pull the trigger every time the light turns green. As in Method #1, the PDT method is typically organized using the pre-test/post-test format, and the results are again scored in terms of RT and accuracy. However, unlike Method #1, the performance of the primary and secondary tasks are typically analyzed together. Learning of the primary task is said to have reached automaticity when both tasks can be performed to a pre-defined level of proficiency with neither task interfering with the other.

Applying the dual-task method may or may not be a challenge, depending on the specific simulation platform that one is using. For example, in a driving simulator, one could mount a box of LED lights in the periphery of the participant's visual field, and attach a micro-switch to the steering wheel to record the learner's response. In a study on driver workload, this method was used to differentiate the performance of novice and expert drivers under different levels of task complexity. Not surprisingly, experts had quicker reaction times and greater accuracy than novices (Patten, Kircher, Ostlund, Nilsson, & Svenson, 2006). The secondary task does not need to be visual, however; it could be an auditory one, such as talking on a cell phone. In a study of distracted driving, novice and expert drivers were required to verbally answer a series of complex math problems using a hands-free cell phone while driving at various speeds. Performance on the primary driving task was measured in terms of steering wheel deviations, with larger numbers indicating poorer performance; performance on the secondary math was calculated as the number of math errors, with larger numbers also indicating poorer performance. Over five days of practice, the participants' steering wheel deviations approached zero for all learner groups. Moreover, the number of math-related errors dropped between 30-50%, depending on age, with the largest reductions occurring among older drivers (Shinar, Tractinski, & Compton, 2005). It is important to note that the choice of secondary tasks can have a profound impact on the measurement of learning. For example, two tasks that use the same sensory channel (e.g., two visual tasks or two auditory tasks) will impose a greater workload than two tasks that use separate sensory channels (e.g., one visual and one auditory task) (Wickens & McCarley, 2008). As a result, comparable levels of performance on secondary tasks that use the same sensory channel provide a much more stringent measure of learning than secondary tasks which use different sensory channels.

Method #3: Overlearned Responses

At the highest levels of expertise, individuals can perform domain-specific tasks virtually error-free and with little conscious attention. This stage of learning is referred to as the "autonomous stage," because performance has become nearly automatic in its execution (Fitts & Posner, 1967). However, because the behavior is automatic, it can also be very difficult to suppress, even when it is situationally inappropriate or counterproductive. For example, expert pilots who transition to a new aircraft sometimes revert to their previous overlearned response (from the prior aircraft) under stress, often with tragic results. This phenomenon is known as "habit lag" (Fitts & Posner, 1967; Toner, Montero, & Moran, 2015). Assessment method #2, described above, is based on the premise that automaticity should enhance performance across a wide range of conditions due to reduced demands on working memory. By contrast, method #3 is based on the premise that automaticity can be assessed by systematically altering the task characteristics, such that the automatic performance would actually result in lower scores. In practice, this method would use the pre-test/post-test assessment paradigm. Prior to training, learners would perform the task two ways: in its standard form, and in a deliberately modified form. At the end of training, learners would again perform the same two tasks. To the extent that learning becomes automatic, the learners' performance on the standard task should increase, and their performance on the deliberately modified task should decrease.

For example, in a study of driving performance, Korteling (1994) had a sample of expert (older) and novice (younger) participants closely follow the car in front of them. The participants performed the task under two different conditions. In the *normal polarity condition*, the gas and brake pedals functioned as normal. In the *reverse polarity condition*, the gas and brake pedals functioned in reverse. For example, pressing the accelerator caused the car to slow, while releasing pressure on the accelerator caused the car to speed up. The criterion measure was how closely participants were able to maintain a constant fixed distance (15 meters) between the two vehicles. Not surprisingly, the reverse polarity condition negatively affected the experts' performance more so than it did the novices' performance. Although it has yet to be used in the published literature, this general approach could easily be applied to the assessment of perceptual skills. Consider the case of a pilot's Heads Up Display (HUD). The HUD uses augmented reality to project critical data onto the external environment – such as the jet's airspeed, altitude, and angle of attack – so that the pilot does not need to shift attention between the cockpit instruments and the out-the-window view. To the extent that the HUD information is consistent with what is actually occurring outside the cockpit, it should enhance the users' performance. Moreover, it should enhance the performance of experts more than novices, because experts can automatically integrate both sets of data – the environmental picture and the HUD data – in real time. However if the HUD information substantially conflicted with the out-the-window image – for example, if there was a significant delay in the rate at which the HUD symbology and data were updated – the result should negatively affect performance for the experts more so than for the novices. The effect would not likely be limited to HUDs, but could be applied to any real-time decision support tool, such as the Improved Fresnel Lens Optical Landing System (IFLOLS) that Navy pilots use to maintain the proper glideslope during carrier landings, and the Global Positioning System (GPS) that Navy surface vessel and submarine navigators use to precision navigate in complex waterways[2].

Method #4: Unobtrusive Physiological and Neurophysiological Measures

The brain bases of System 1 (automatic processing) and System 2 (deliberate processing), from their knowledge structures to their supporting cognitive processes, likely differ in both location and activation patterns. As a result, expertise should be reflected in neurophysiological recordings, such as brain wave activity collected with electroencephalograph (EEG) sensors or in cerebral blood oxygenation patterns collected with functional Near Infrared (fNIR) spectroscopy sensors. Generally speaking, these patterns of neurophysiological changes are initially revealed by comparing known groups of experts and novices during controlled task performance. They are later verified by following a cohort of learners over time as they become more expert-like in their domain. For example, in an EEG study of perceptual expertise (category learning), the peaks of early brain wave components were more negative when bird or dog experts viewed pictures of animals within their domain of expertise than when compared to other types of animals (Tanaka & Curran, 2001). In other words, functional electrical brain components differed according to expertise in category learning, in this case, categories of animals. In a second study, Scott and colleagues (2006) showed that object categories can be learned and, furthermore, that the same early brain wave components mentioned above reflected this change from pre- to post-training. Similarly, in a fNIR study of learning to use a new air traffic control management tool, Harrison and colleagues (2014) recruited twelve certified Air Traffic Controllers (ATCs), to perform nine learning trials over the course of three days (three trials per day). The results showed a significant decrease in prefrontal cortex blood oxygenation (an indicator of cognitive workload) after the first day of instruction, and the results remained stable during the second and third days. In other words, the learners needed only a full day of instruction to learn the new technology, after which time their workload dropped precipitously.

Both EEG and fNIR systems both involve a cap or pad, containing sensors (electrodes in the case of EEG, photodetectors in the case of fNIR), that is worn on the head. The sensors are connected to devices that amplify, record, and/or pre-process the signals. The signals can then be pushed to a computer for additional processing or analysis. In practice, unobtrusive measurement methods are attached to the participant prior to the training exercise, calibrated for accuracy, and worn throughout the entire training event. It requires only minimal training to physically attach the sensors to the learner, and to connect system components to one another. However, the interpretation of the resulting signal data is very complex. Neurophysiological sensors are capable of recording data at much higher rates than behavioral approaches. For example, an EEG system can sample multiple variables at approximately 500 times per second; by comparison, a typical 2-second long decision-making trial results in only one RT score and accuracy score each. Additionally, neurophysiological methods tend to cover large areas of interest and record slightly different activity per location. For example, a 64 sensor EEG system covers 64 unique spatial locations across the scalp, which translates into many functionally-specific cortical regions in the brain. Therefore, a good understanding of human physiology and the published research is useful when determining how to prune, process,

[2] While these systems are highly reliable, thereby reducing the probability that HUD, IFLOLS, and GPS degradations would occur in real life, 100% real-world fidelity is not necessary when it comes to training and assessing skilled performance There are many examples where training deliberately differs from real world conditions. For example, some lifelike patient manikins allow users to hide different "layers" of body tissue (e.g., muscle and bone) so that they can get an unobstructed view of the body's internal organs. Other simulators use visual or auditory cues during CPR training to help learners determine whether or not they are applying the appropriate amount of force during chest compressions. Obviously, neither of these cues exist in the real world, but are nonetheless useful during skill acquisition.

segment, and examine the data (Luck, 2005; Picton, Bentin, Berg, Donchin, Hillyard et al., 2000). Finally, a common axiom in this field states, “there is no better data than cleanly-collected data.” In other words, it is much more difficult to clean-up poorly-recorded data (after the fact) than it is to take an extra 15-20 minutes to improve the sensors’ connections to the learner’s scalp (prior to and during the experiment). In addition, ideal recording conditions – such as low humidity, depending on the specific neurophysiological assessment technique – greatly improve one’s chances of recording a better ratio of physiological signal relative to noise. At the same time, these conditions may impose practical limitations on their use outside of climate-controlled laboratory settings.

MEASUREMENT METHODS INSPIRED BY THE GENERALIZED MODEL OF SKILL ACQUISITION

In this section, we review five measurement methods that capitalize on well-known qualitative differences as one progresses from novice to competent, proficient, and eventually expert. For example, with increased expertise, knowledge is better organized; decision-making becomes more goal-directed; and performance becomes proceduralized.

Method #5: Structural Knowledge Assessment

As learners becomes more expert-like, they do not simply accumulate more and more facts about the content domain. Rather, they organize these facts based on meaningful relationships among important domain-specific concepts. The way that these concepts are organized is known as a “knowledge structure” (Goldsmith & Kraiger, 1997). At the most extreme level, novices only understand the task or domain in the abstract. As a result, their knowledge structures are poorly-formed. For example, important concepts may only be weakly linked to one another, while others may be linked by only a single pathway. By contrast, experts’ knowledge structures tend to be much more rich and nuanced. They are much more likely to have strong linkages among important domain concepts, and these concepts are usually linked together via multiple pathways, thereby facilitating rapid information retrieval. The differences in how novices and experts organize their knowledge can be measured quantitatively and displayed graphically. Specifically, factual concepts can be represented as nodes on a graph, and the relationships among them can be expressed as linkages among the nodes. The strength of those linkages is a joint function of their closeness (proximity) and number of pathways (complexity). Knowledge structures can be scored in two ways: internal consistency (the H index), which measures the extent to which the structure is mathematically coherent; and similarity of an individual’s knowledge structure to that of a knowledgeable other (the C index), such as an instructor. Both scores range from 0.0 to +1.0, with larger numbers representing better knowledge organization.

The structural knowledge assessment method (Goldsmith & Kraiger, 1997) is typically performed using the pre-test/post-test assessment paradigm. Prior to the start of training, the instructor uses a tool such as *Pathfinder* to display key domain concepts on a computer screen. For each unique pair of concepts, the learner rates its similarity using a 5-point Likert scale. After rating all unique item pairs, the data are then converted to a proximity matrix for statistical analysis (Schvaneveldt, 1990). At the end of training, the assessment process is repeated. Finally, difference scores are calculated between the pre-training (baseline) and post-training (learning) assessments. The amount of learning is calculated as increases in internal consistency (the H index), and/or as improved similarity between the learners’ and instructor’s knowledge structures (the C index). This method has previously been used to assess the learning of teamwork skills among Navy pilots (Stout, Salas, & Kraiger, 1997). In that study, approximately half of the learners were randomly assigned to an experimental training condition; the other half were randomly assigned to a control condition. At the end of training, learners in the experimental condition had significantly higher scores on the C index than did those in the control group, thereby demonstrating that their knowledge structure became more similar that of the instructors. The learners in the experimental condition also scored higher on the H index than did learners in the control condition, thereby demonstrating that their knowledge structure was also more internally consistent. Recently, a more efficient way has been developed to assess structural knowledge networks. Instead of presenting each unique pair of concepts on separate screens, this new method uses a target-style graphical user interface that allows users to rate the level of similarity among several concepts by dragging them onto a target, with the inner concentric rings representing greater similarity than the outer rings. While the target method does not take less time than the traditional pairwise rating method, the results are comparable to the pairwise method. In addition, many participants prefer the target-based assessment approach over the pairwise method (Tossel, Smith, & Schvaneveldt, 2009).

Method #6: Comparative Learning Curves

One of the hallmarks of expertise is the ability to demonstrate consistently superior levels of performance vis-à-vis one's peers, and to do so even under adverse conditions. For example, an expert marksman should be able to consistently hit the target over repeated trials. Moreover, he or she should be able to hit the target from various distances, under conditions of reduced visibility, and even under heavy crosswinds. Experts' superior performance is the result of sustained, deliberate practice – a systematic training regimen that involves strategic goal-setting, focused skills practice, process-based feedback, and deliberate self-reflection – over time. Some researchers have suggested that it takes 4-5 hours of deliberate practice every day for up to 10 years in order to achieve expert-like levels of proficiency (Ericsson, Krampfe, & Tesch-Romer, 1993). These research findings have three profound implications for the measurement of learning, particularly with regard to training programs that involve short time periods ranging from hours to weeks. First, at baseline prior to the start of training, the experts should outperform the learners by a considerable margin. Second, since experts have developed and honed their skills over the course of many years, their performance should not appreciably improve during the brief time period. Third and finally, the experts' performance should serve as a benchmark for systematically comparing the extent to which novices have learned.

Applying this assessment method is relatively simple in practice. One would first need to recruit a small sample of experts to serve as the referent comparison group. Immediately prior to training, the proficiency of both groups (learners and experts) would be baselined using a series of standardized tasks or scenarios in the simulator. Both groups would then undergo the training regimen. At several pre-defined points in the curriculum (and again at the end of training), both groups would again perform the standardized tasks or scenarios to quantify the amount of learning. Finally, the performance of each group would be averaged and plotted separately as a function of time. Generally speaking, the experts' performance should start out higher than the novices and remain consistent over time, since they are unlikely to learn over the short term. By comparison, the novices' performance should start out lower than the experts, and should steadily improve over time. The results could be analyzed using statistical techniques such as Repeated Measures Analysis of Variance (RMANOVA) to determine if the results are statistically significant. For example, in a study on the acquisition of endoscopic surgical skills, a sample of novices (medical students) and experts (gynecologists) each performed nine separate learning trials. Eight separate criterion measures were collected, including time to completion, efficiency of motion, and unintentional tissue damage, among others. Not surprisingly, at the start of training, the experts outperformed the novices on all eight criterion measures. However, over the course of time, novices reached expert-like levels of proficiency on six of the eight criterion measures (Janse, Goedegeburre, Veersema, Broekmans, & Schreuder, 2013).

Method #7: Detection of Subtle Cues and Rank-Ordering of Priorities

With increased experience, novice learners reach the competent stage. At this stage, they are able to identify situations that clearly call for one decision choice versus another. For example, they can identify examples of clearly hostile and non-hostile enemy behaviors. However, their identification of subtle situational cues requires assistance from the instructor or mentor. In terms of their decision-making, competent learners rely entirely on instructor-provided guidelines or heuristics. However, since they fail to detect subtle situational cues, they treat all such guidelines as being equally important, regardless of the unique situational characteristics that would otherwise tend to favor one guideline more so than the others (Dreyfus & Dreyfus, 1980). By comparison, proficient learners are better able to pick up on subtle contextual cues. Moreover, since their behavior is much more goal-directed, they can use these cues to help them decide which decision-making guidelines are most relevant to the particular situation.

Historically, Dreyfus-inspired assessment methods for differentiating between competent and proficient learners have been limited to observer-based ratings (Bondy, 1983; Phillips, Shafer, Ross, Cox, & Shadrick, 2006). However, they could also be computerized. For example, with regard to perceptual skills, one way to assess learning might be to develop a set of still images – such as battlefield photographs or computer-generated images developed using a simulator such as *Virtual Battle Space* – each of which contains several subtle threat cues. The images can then be presented randomly on a computer screen. Using a mouse, learners would then mark the location of these threats. Various types of criterion measures could potentially be collected, such as response time (RT), decision accuracy, priority (as determined by a panel of Subject Matter Experts), and the like. In a recent study of visual threat detection skills, Zimmerman and colleagues (2013) developed a series of computer-based assessment tasks. One of these methods, the *Limited Threat Search* task, presented learners with a series of still images. Learners were then instructed to locate potential threats in each image by clicking on their locations. Criterion measures included decision accuracy and RT (in milliseconds). Another assessment method, the *Decision Making Exercise*, presented the learners with a brief scenario that described a challenging patrol; they were also given a photo that helped to explain the scenario text. They were then given two minutes to study the materials. Finally, they were then asked to describe the threats presented in the scenario, the meaning of those threats, and the specific threat-related cues that informed their decisions. The learners' text responses were then content coded using a Dreyfus model-inspired scoring system, with performance scored as being "novice," "advanced beginner," "competent," "proficient," or "expert." Qualitative results suggested that differing levels of expertise were reflected in the quality of the responses. However, the study lacked statistical power, and failed to reveal any statistically significant effects. Nevertheless, the method remains promising.

Method #8: Task Compilation and Fluidity of Movement

There are several predictable changes in motor task-related performance as one progresses across the various stages of skill acquisition (Kraiger, Ford, & Salas, 1993). For example, novices generally perform the task exactly as it is described in the textbook or by the instructor. They frequently hesitate between performing the steps of a multi-step procedure. When performing the task without any external guidance or cues, they may occasionally perform redundant steps, forget key steps, or perform steps in the wrong order. As a general rule, their performance tends to be slow, effortful, and jerky. Finally, they have a tendency to misjudge the amount of force required to perform the task. As a result, they may have to compensate for prior actions (e.g., abruptly steering left because they previously steered too hard right) and/or cause accidental damage (e.g., hitting objects that should have been avoided). By comparison, competent learners' motor performance should be somewhat faster and less error-prone. In addition, their overall time to completion, as well as time to completion for specific task steps, should decrease. They should perform task steps in a more logical sequence, and there should be a reduction in the number of redundant steps. As learners reach the proficient stage, their motor task performance should become smoother, more accurate, and more efficient. They may begin to intentionally deviate from textbook descriptions, for example, by performing certain steps in parallel, by combining and integrating steps to achieve efficiency, or by adapting steps and procedures based on unique environmental conditions. However, unlike true experts and masters, they are unlikely to develop entirely new ways or methods of performing the task (Kraiger, Ford, & Salas, 1993).

Although conceptually simple, these types of learning measures can be very difficult to implement in practice. This is due to the fact that many of the variables measured by a simulator often do not directly map onto the specific types of behaviors that the trainer is interested in measuring. As a result, multiple data fields often must be combined, transformed, and aggregated to form meaningful performance measures. Moreover, the performance measures need to then be compared across time and/or training scenarios to measure learning. For example, in a study on the effectiveness of simulation-based training for improving laparoscopic surgical suturing skills, a sample of novices (first year residents) and experts (surgeons) each performed 10 learning trials. Five different criterion measures were collected, including time to completion, number of errors, gesture proficiency, hand movement smoothness, and tool movement smoothness. Certain measures – such as the number of errors – were collected directly from the simulator, and were measured based on deviations from the ideal needle placement, penetration depth, and application of force, as determined by Subject Matter Experts. Other measures – such as gesture proficiency – were recorded using motion-sensing gloves and micro-sensors that were affixed to the surgical instruments. However, the sensors could only measure a limited number of basic movements, such as up, down, left, right, in, out, rotation clockwise, rotation counterclockwise, and grasping. Therefore, algorithmic models of expert task performance first had be developed based on the quality, combination, sequence of basic movements required to complete the entire suturing task. Later the novices' performance were then compared to the experts' performance, and scaled from 0-10, with 0 representing the lowest level of similarity with expert performance and 10 representing the highest level of similarity (Hamilton, Kahol, Vankipuram, Ashby, Notricia, & Ferrara, 2011). Finally, performance was plotted as a function of practice trials to demonstrate learning. Obviously, once the algorithms were developed, they could be re-used to measure suturing skills. However, a non-trivial amount of effort went into their development and validation prior to use.

Method #9 Anticipatory Behaviors

In comparison to novices, experts have superior perceptual-cognitive (Klein & Hoffman, 1992) and perceptual-motor (Suss & Ward, 2015) skills. Based on their extensive repertoire of prior experiences, experts can quickly identity examples of “typical” situations, attend to the most relevant environmental cues, and rapidly respond by selecting actions that have worked well in similar situations. Moreover, through the process of mental simulation, experts are able to visualize how a situation likely arrived at its current state, as well as how it will likely unfold in the near future (Klein & Hoffman, 1992). For example, in baseball expert batters can anticipate the ball’s likely trajectory and vertical distance above the ground when it crosses home plate – based entirely on cues that occur before the pitcher has even released the ball.

There are several standard techniques for measuring anticipatory skills in perceptual-motor tasks. For example, in the temporal occlusion method, a first-person video is presented to the learner. In the case of a baseball batter, the video would depict a pitcher winding up to throw the ball. At a critical moment in the wind-up, the screen is blanked, and the learner is required to swing the bat in response to the ball’s anticipated trajectory. Similarly, in the spatial occlusion method, the participant views an event sequence, such as a group of soccer players maneuvering the ball downfield. At a critical moment in the action, one or more players is then digitally blurred (or removed entirely) from the video. The participant then needs to quickly determine where the ball will likely travel (Suss & Ward, 2014).

Historically, these methods have been used largely in the domain of sports psychology. However, a recent I/ITSEC paper by Stacy and colleagues (Stacy, Beaubien, Wiggins, Walwanis, & Bolton, 2014) showed how this assessment technique can be used to assess the visual perceptual skills that are required for carrier landings. Specifically, expert and novice Navy pilots were shown a series of brief 8-second videos, each of which depicted a portion of the final approach pattern. At the end of each video clip, the screen was immediately masked, and the participants needed to rapidly determine whether that pilot would next need to make a standard (small) or major (large) correction. As expected, the results clearly showed a main effect of expertise on decision-making performance, with experts having higher levels of accuracy and quicker reaction times than novices. Curiously, the results did not demonstrate an effect of learning over time. However, the experimental manipulation (simulator motion cues) that were present during training were not present during the PC-based transfer task, which may have influenced the data. The paper also describes several standard practices for processing and analyzing RT data. For example, since reaction times are not normally distributed, the data must be converted to logarithms prior to statistical analysis. In addition, Suss and Ward (2013) provide a relatively simple way to help identify video clips that adequately discriminate between experts and novices. They also provide guidance on identifying the optimal time point where the video should be occluded.

CONCLUSIONS

This paper summarized nine theoretically-derived methods for measuring learning. Many of these methods can be implemented using relatively inexpensive commercial off-the-shelf (COTS) hardware and/or software. Moreover, with the exception of a small handful of methods, most can be applied by training professionals with a little practice. After finishing this paper, some readers might wonder “*Why not simply use a pre/post knowledge test to measure learning?”* While knowledge tests can be developed quickly and inexpensively, they suffer from several critical drawbacks. First, this approach is designed to assess factual knowledge, rather than proceduralized skill. Second, a great deal of expertise lies beyond conscious awareness. Therefore, experts often find it difficult to clearly articulate why they perform the task as they do. Third and finally, good knowledge tests can be very difficult to develop, and often require multiple rounds of pilot testing to ensure that there is sufficient variability in responding.

Alternatively, some readers might wonder “*Why not simply use a pre/post assessment in the simulator to measure skills improvement?”* This is a very common technique. Unfortunately, if not applied properly, it can provide very flawed information about learning. At the most basic level, every simulator is an abstraction of reality. Therefore, learners often behave very differently in the simulator than they do in the real world environment (Fowlkes, Sheehan, Milham, Pagan, & Ashlock, 2011). As a result, unless one takes great care in the design of the study – for example by including experts as a comparison group, and/or by using transfer tasks – it can be difficult to rule out the possibility that the participants merely “learned to fly the simulator,” rather than learned the critical skills.

Many readers may still be unconvinced, and might wonder “*Why the need for such a diversity of measurement methods?”* As noted previously, these measurement methods are rooted in well-respected theories of skill acquisition and expertise (Dreyfus & Dreyfus, 1980; Kraiger, Ford, & Salas, 1993; Anderson, 1983; Laird, 2012) that hypothesize well-defined cognitive, physiological, and behavioral changes as individuals acquire increasingly greater levels of skill over time. These methods are useful in that they can help training professionals focus their efforts on what to measure, how to measure it, and when to measure it. In essence, the underlying theory provides a roadmap to help training professionals ensure that they are looking in the right places for changes in learning, rather than applying measurement methods haphazardly.

Finally, some readers may be tempted to just throw their hands in the air and say *"This is way too complicated."* The measurement of learning is complicated, but not overly so. As a reminder to the reader, learning represents a relatively permanent change in knowledge, skills, or understanding across tasks, across time, and/or across environments. As a result, measures of learning should incorporate different types of tasks, different time periods, and/or different environmental conditions in order to assess skill retention and transfer. Obviously, nobody could possibly incorporate all of these measurement methods into their training-related efforts. Doing so would be cost and time prohibitive. However, to the extent that this paper has caused the reader to at least re-examine his or her own perspectives on how to measure learning – even if they don't necessarily adopt these methods *per se* – it will have been a success.

ACKNOWLEDGEMENTS

Funding for this paper was provided by the Office of Naval Research to Aptima, Inc., NAWCTSD, and NAWCAD (Contract No. N00014-12-C-0058). The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Office of Naval Research (ONR), the Naval Air Warfare Center (NAWC), or any other Department of Defense (DoD) agency, unless stated in official directives.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Beaubien, J., Stacy, E., & Wiggins, S., Keeney, M., Bolton, A., Grubb, J., Walwanis, M., Priest, H., & Riddle, C. (2015). Differentiating measures of learning (MOL) from measures of performance (MOP) during aircraft carrier landing practice. Paper No. 15210. In *Proceedings of the 2015 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Arlington, VA: National Training and Simulation Association.

Bondy, K., (1983). Criterion-referenced definitions for rating scales in clinical evaluation. *Journal of Nursing Education, 22*, 376-382.

Christina, R., & Bjork, R. (1991). Optimizing long-term retention and transfer. In D. Druckman & R. Bjork (Eds.), *In the mind's eye: Enhancing human performance* (pp. 23-56). Washington, DC: National Academy Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Dreyfus, S., & Dreyfus, H. (1980). *A five-stage model of the mental activities involved in direct skill acquisition*. Report No. ORC 80-2. Berkeley, CA: Operations Research Center, University of California Berkeley.

Evans, J. St. B.T., & Stanovich, K.E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives On Psychological Science, 8*, 223–241.

Ericsson, K., Krampe, R., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363-406.

Fairclough, S., Venables, L., & Tattersall, A. (2005). The influence of task demand and learning on the psychophysiological response. *International Journal of Psychophysiology, 56*, 171-184.

Fitts, P., & Posner, M. (1967). *Human performance*. Oxford, England: Brooks/Cole.

Fowlkes, J., Sheehan, J., Milham, L., Pagan, J., & Ashlock, D. (2011). *Aircraft carrier approach and landing fidelity analysis (ALFA)*. Technical Report No. NAWCTSD-TR-2012-0001. Orlando, FL: Naval Air Warfare Center Training Systems Division.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*, 236-243

Goldsmith, T., & Kraiger, K. (1997). Structural knowledge assessment in training evaluation. In J. Ford, S. Kozlowski, K. Kraiger, E. Salas, & M. Teachout (Eds.), *Improving training effectiveness in work organizations (pp. 73-96)*. Mahwah, NJ: Erlbaum.

Janse, J., Goedegeburre, R., Veersema, S., Broekmans, F., & Schreuder, H. (2013). Hysteroscopic sterilization using a virtual reality simulator: Assessment of learning curve. *The Journal of Minimally Invasive Gynecology, 20*, 775-782.

Hamilton, J., Kahol, K., Vankipuram, M., Ashby, A., Notrica, D., & Ferrara, J. (2011). Toward effective pediatric minimally invasive surgical simulation. *Journal of Pediatric Surgery, 46*, 138-144.

Harrison, J., Izzetoglu, K., Ayaz, H., Willems, B., Hah, S., Ahlstrom, U., Woo, H., Shewokis, P., Bunce, S., & Onaral, B. (2014). Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy. *IEEE Transactions on Human-Machine Systems, 44*, 429-440.

Hull, C. (1943). *Principles of behavior*. New York: Appleton-Century-Crofts.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*, 515-526.

Klein, G., & Hoffman, R. (1992). Seeing the invisible: Perceptual-cognitive aspects of expertise. In M. Rabinowitz (Ed.), *Cognitive science foundations of instruction (pp. 203-226)*. Mahwah, NJ: Erlbaum.

Korteling, J. (1994). Effects of aging, skill modification, and demand alternation on multiple task performance. *Human Factors, 36*, 27-43.

Kraiger, K., Ford, J., & Salas, E. (1993). Application of cognitive, skill-based and affective theories of learning to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311-328.

Laird, John E. (2012). *The Soar Cognitive Architecture*. MIT Press.

Luce, R. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Luck, S. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*, 314-324.

Martens, M., & van Winsum, W. (2000). *Measuring distraction: the peripheral detection task*. Soesterberg, Netherlands: TNO Human Factors. http://www-nrd.nhtsa.dot.gov/departments/Human%20Factors/driver-distraction/PDF/34.PDF.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson (Ed.), *Cognitive skills and their acquisition (pp. 1-51)*. Hillsdale, NJ: Erlbaum.

Patten, C., Kircher, A., Ostlund, J., Nilsson, L., & Svenson, O. (2006). Driver experience and cognitive workload in different traffic environments. *Accident Analysis and Prevention, 38*, 887-894.

Phillips, J., Shafer, J., Ross, K., Cox, D., & Shadrick, S. (2006). *Behaviorally Anchored Rating Scales for the Assessment of Tactical Thinking Mental Models*. Research Report No. 1854. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences.

Picton, T., Bentin, S., Berg, P., Donchin, E., Hillyard, S., Johnson Jr., R., Miller, G., Ritter, W., Ruchkin, D., Rugg, M., & Taylor, M. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology, 37*, 127–152.

Pierce, J. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroscience, 2*. http://dx.doi.org/10.3389/neuro.11.010.2008

Schvaneveldt, R. W. (1990). *Pathfinder associative networks*. Norwood, NJ: Ablex.

Servan-Schreiber, E. & Anderson, J. R. (1990). Chunking as a mechanism of implicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592-608.

Shinar, D., Tractinsky, N., & Compton, R. (2005). Effects of practice, age, and task demands, on interference from a phone task while driving. *Accident Analysis and Prevention, 37*, 315-326.

Stacy, E., Beaubien, J., Wiggins, S., Walwanis, M., & Bolton, A. (2014). Using temporal occlusion to assess carrier landing skills. In *Proceedings of the 2014 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Paper No. 14171. Arlington, VA: National Training and Simulation Association.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

Suss, J., & Ward, P. (2015). Predicting the future in perceptual-motor domains: Perceptual anticipation, option generation, and expertise. In J. Szalma, M. Scerbo, R. Parasuraman, P. Hancock, & R. Hoffman (Eds.), *The Cambridge handbook of applied perception research (pp. 951–976)*. New York, NY: Cambridge University Press.

Suss, J., & Ward, P. (2013). Investigating perceptual anticipation in a naturalistic task using a temporal occlusion paradigm: A method for determining optimal occlusion points. In *Proceedings of the 57th Annual Meeting of the Human Factors and Ergonomics Society (pp. 304-308)*. Santa Monica, CA: HFES.

Toner, J., Montero, B., & Moran, A. (2015). The perils of automaticity. *Review of General Psychology, 19*, 431-442.

Tossel, C., Smith, B., & Schvaneveldt, R. (2009). The influence of rating method on knowledge structures. In *Proceedings of the 53rd Annual Meeting of the Human Factors and Ergonomics Society (pp. 1893-1897)*. Santa Monica, CA: HFES.

Tanaka, J. W. & Curran, T. (2001). A neural basis for expert object recognition. *Psychological Science*, *12*, 43-47.

Wickens, C., & McCarley, J. (2008). *Applied attention theory*. Boca Raton, FL: CRC Press.

Wright, R., Thompson, W., Ganis, G., Newcombe, N., & Kosslyn, S. (2008). Training generalized spatial skills. *Psychonomic Bulletin & Review*, *15*, 763-771.

Zimmerman, L., Liens, D., Marcon, J., Pearlman, J., Singer, J., Mueller, R., & Vowels, C. (2013). *Improving Visual Threat Detection: Research to Validate the Threat Detection Skills Trainer*. Research Report No. 1329. Ft. Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.