Differentiating Measures of Learning (MOL) from Measures of Performance (MOP) During Aircraft Carrier Landing Practice

Jeffrey M. Beaubien[1], E. Webb Stacy[1], Sterling L. Wiggins[2], Michael J. Keeney[3]
Aptima, Inc.
Woburn, MA[1], Fairborn, OH[2], Washington, DC[3]
jbeaubien@aptima.com , wstacy@aptima.com , swiggins@aptima.com , mkeeney@aptima.com

Amy E. Bolton
Office of Naval Research, Code 34
Arlington, VA
amy.bolton@navy.mil

LCDR Jefferson D. Grubb
Naval Aviation System Program Office (PMA-205)
Patuxent River, MD
jeff.grubb@navy.mil

Melissa M. Walwanis, Heather Priest
Naval Air Warfare Center Training Systems Division
Orlando, FL
melissa.walwanis@navy.mil , heather.priest@navy.mil

Christian S. Riddle
Naval Air Systems Command, Manned Flight Simulator
Patuxent River, MD
christian.riddle@navy.mil

ABSTRACT

Measures of performance collected during initial skill acquisition can be misleading indicators of long-term retention or transfer (Soderstrom & Bjork, 2013). For example, previous research demonstrates that learning can occur in the absence of visible performance gains, and temporary performance gains can occur in the absence of long-term retention or transfer (Singer & Edmondson, 2006; Soderstrom & Bjork, 2013). Therefore, it is critical that authors clearly differentiate between Measures of Learning (MOLs) and Measures of Performance (MOPs) in their research. While this distinction was frequently made in the psychological literature until the 1950's, it has been somewhat forgotten since then (Schmidt & Bjork, 1992). As part of a larger study on the effects of simulator cue fidelity on aircraft carrier landing skills, we collected both MOLs and MOPs. The sample included 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. 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. The data were then aggregated across all 24 passes. In contrast, MOLs were operationalized as changes in performance over time. The two sets of analyses – learning vs. performance – provide very different interpretations of the data. In this paper, we describe the conceptual differences between MOLs and MOPs; show how the choice of analysis can have profound implications for interpreting the results; and provide the reader with actionable guidelines that they can use in their own work to better differentiate learning from performance.

ABOUT THE AUTHORS

Jeffrey M. Beaubien, Ph.D. is a Principal Scientist in the Advanced Cognitive Training Systems division at Aptima, Inc. For the past 15 years, his work has focused on training and performance assessment. His work has been sponsored by the U.S. Navy, the U.S. Army, the U.S. Air Force, the Federal Aviation Administration, and the Telemedicine and Advanced Research Technologies Center, among others. Dr. Beaubien received a Ph.D. in Industrial and Organizational Psychology from George Mason University, an 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 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 Program Chair for the Society for Behavior Representation in Modeling and Simulation.

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 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.

Michael J. Keeney, Ph.D. is a Senior Scientist in the Performance Assessment and Augmentation Division of Aptima, Inc. He brings over 18 years of experience in the areas of occupational analysis, personnel selection, and training. Dr. Keeney holds a Ph.D. in Industrial-Organizational Psychology from the University of Akron, a M.A. in Psychology from the University of Akron, a B.A. in Psychology from the University of Maryland, and an A.A. in Technical Management from the Community College of the Air Force.

Amy E. Bolton, Ph.D. is a Program Officer at the Office of Naval Research (ONR). She manages several programs within the Capable Manpower Future Naval Capability. Capable Manpower is a multi-million dollar per year initiative that addresses manpower, personnel, training, and human system design Science and Technology (S&T) challenges for the Navy and Marine Corps. Dr. Bolton's research interests include adaptive training; human behavior modeling; human system design; and Live, Virtual, and Constructive training. She holds a Ph.D. in Applied Experimental and Human Factors Psychology from the University of Central Florida.

LCDR Jefferson D. Grubb, Ph.D. is a Naval Aerospace Experimental Psychologist who currently is the lead for Air Warfare Training Development within PMA-205, the NAVAIR Program Office for Aviation Training. He has previously served as the Military Deputy for Research and Technology at NAWCTSD, head of the Biostatistics Division at the Naval Aerospace Medical Institute, and as a project officer in the NAVAIR Human Systems Engineering Department. He holds a Ph.D. in Developmental Cognitive Neuroscience from the University of Denver, and a B.S. in Psychology from the University of Alaska Fairbanks.

Melissa M. Walwanis, M.S. is a Senior Lead Research Psychologist at NAWCTSD. She has an integrated research program devoted to transitioning state-of-the-art products to enhance the training and operational capabilities of the nation's Warfighters. Her research portfolio addresses large scale distributed training; scenario authoring; performance assessment; training fidelity needs analysis; and Live, Virtual, and Constructive (LVC) training. She has an M.S. in Industrial and Organizational Psychology from the University of Central Florida, and is pursuing a Ph.D. in Industrial and Organizational Psychology at the Florida Institute of Technology.

Heather Priest, Ph.D. is a Senior Research Psychologist at the Naval Air Warfare Center Training Systems Division (NAWCTSD) following six years at the U.S. Army Research Institute (ARI) Technology-Based Training Research Unit in Orlando. At NAWCTSD, her research revolves around fidelity and realism of semi-automated forces (SAF); Live, Virtual, and Constructive training; and adaptive training. Dr. Priest received a Ph.D. in Applied Experimental and Human Factors Psychology from the University of Central Florida, and an M.A. in Experimental Psychology from Mississippi State University.

Christian S. Riddle, B.S. is the Laboratory Architect at Manned Flight Simulator (MFS), which is located at Naval Air Station Patuxent River. As the lead technical advisor for the MFS lab facility, Mr. Riddle is responsible for identifying, defining, planning, prioritizing, developing, and managing technical solutions for the simulation hardware and software systems. He received a B.S. in Computer Engineering from Ohio University.

Differentiating Measures of Learning (MOL) from Measures of Performance (MOP) During Aircraft Carrier Landing Practice

Jeffrey M. Beaubien[1], E. Webb Stacy[1], Sterling M. Wiggins[2], Michael J. Keeney[3]
Aptima, Inc.
Woburn, MA[1], Fairborn, OH[2], Washington, DC[3]
jbeaubien@aptima.com , wstacy@aptima.com , swiggins@aptima.com , mkeeney@aptima.com

Amy E. Bolton
Office of Naval Research, Code 34
Arlington, VA
amy.bolton@navy.mil

LCDR Jefferson D. Grubb
Naval Aviation System Program Office (PMA-205)
Patuxent River, MD
jeff.grubb@navy.mil

Melissa M. Walwanis, Heather Priest
Naval Air Warfare Center Training Systems Division
Orlando, FL
melissa.walwanis@navy.mil , heather.priest@navy.mil

Christian S. Riddle
Naval Air Systems Command, Manned Flight Simulator
Patuxent River, MD
christian.riddle@navy.mil

BACKGROUND

Like every Department of Defense organization, the Naval aviation enterprise seeks to identify the appropriate mix of live, virtual, and constructive (LVC) training methods that will ensure the highest levels of readiness while maximizing return on investment. Historically, the percentage of flight training conducted in the simulator has been a small fraction of the pilots' total training time. However that percentage is expected to increase moving forward. For example, during Fiscal Year (FY) 2010, Navy F/A-18 pilots were required to complete 81.9% of their training using live fly methods; 12.5% using a combination of live fly and simulator-based methods; and 5.6% exclusively in the simulator. Looking toward FY 2015 and beyond, those percentages are expected to become 68.1%, 23.6%, and 8.3%, respectively (Fuller, 2011; Government Accounting Office, 2012). In essence, there will be less training time conducted exclusively using live fly methods, and more training time using simulation-based methods, either alone or in combination with live flight. This shift in the mix of LVC training resources is not unique to the F/A-18 community. It is being applied to several other Navy airframes, including the EA-18G, the P-3C, the MH-60S, and the MH60-R (Fuller, 2011; Government Accounting Office, 2012).

While Naval aviation makes significant use of simulators for skills practice, graded training events are still conducted using live fly methods, such as Field Carrier Landing Practice (FCLP; Government Accounting Office, 2012). This is due to limitations in the simulator fidelity cues that are used to train and elicit the critical piloting skills and behaviors. Historically, subtle cues such as the sea state and the ship's wake have been poorly represented in flight simulators (Fowlkes, Sheehan, Milham, Pagan, & Ashlock, 2011; Wiggins, 2013). As a result, it can be difficult for pilots to accurately judge the ship's speed, the atmospheric wind conditions, and the surrounding sea state. Similarly, the Improved Fresnel Lens Optical Landing System (IFLOLS), a key landing aid that helps the pilots to stay on the correct glide path, is less responsive in the simulator than at the ship. Moreover, the size of the IFLOLS has been artificially enlarged vis-à-vis the ship model so that it can be seen at a distance. As a result, it looks artificially large when viewed up close during the final critical seconds of approach and landing (Wiggins, 2013). Finally, some pilots have reported that there is too much lag between when a pilot manipulates the flight controls and when the simulator's motion platform responds to those inputs (Wiggins, 2013). Not surprisingly,

pilots often report that they fly differently in the simulator than they do at the ship (Fowlkes et al., 2011), which raises concerns about the potential for negative transfer. Because of these and other concerns, Navy pilots do not currently receive training and readiness (T&R) credit in the simulator for the final approach and landing phases of Carrier Qualification (CQ) training; such credit must be earned using live fly methods such as FCLP. To the extent that these simulator cue fidelity limitations can be resolved, it may be possible to use simulators for graded CQ training events, thereby freeing up live fly resources for other skills that cannot be trained in the simulator.

The study described in this paper explored the effects of targeted simulator cue fidelity improvements on pilot learning during the final approach and landing phases of CQ skills training. Based on prior work (Fowlkes et al., 2011; Wiggins, 2013), we identified and prioritized a list of critical simulator enhancements that we believed would improve learning. These included: improved visual resolution and scene generation effects (i.e., bow wake, hull wake, and sea state); improved IFLOLS representation (i.e., improved IFLOLS responsiveness to flight control inputs, more realistic light intensity and shape, and correct size vis-à-vis the ship model); and improved motion cues (i.e., better synchronization among flight control inputs, visual cues, and motion cues). We then developed and refined these enhancements during a series of iterative engineering tests with recently-retired Navy Landing Signals Officers (LSOs) and test pilots to ensure that the upgrades were working as intended. Finally, we collected experimental data from fifteen active duty Navy F/A-18 pilots during the Spring and Fall of 2014. Both MOLs and MOPs were computed from the experimental data.

LEARNING VS. PERFORMANCE

Before discussing our method and findings, it is important to conceptually distinguish learning from performance. While they are related concepts, they are not synonymous. Learning is frequently defined as a relatively permanent change in knowledge, understanding, skills, and/or the capacity to respond (Christina & Bjork, 1991; Soderstrom & Bjork, 2013). Over the years, learning has been described in terms of "habit strength," (Hull, 1943), "reflex reserve," (Skinner, 1938), or "storage strength" (Bjork & Bjork, 1992). It is frequently operationalized in terms of durability, which is the resistance to forgetting (i.e., forgetting curves) or the rapidity of relearning after a period of disuse (i.e., relearning curves). It can also be operationalized as transfer, which is the extent to which knowledge and skills generalize to new tasks, new situations, or both. Learning is not something that can be easily observed. Rather, it must be inferred from patterns of behavior that are measured over time, across contexts, or both (Christina & Bjork, 1991; Soderstrom & Bjork, 2013). By contrast, performance refers to behavioral responses that can be directly observed during and after skill acquisition (Christina & Bjork, 1991; Soderstrom & Bjork, 2013). It is frequently operationalized as the quality, probability, rate, or latency of a particular response. Over the years, performance has been variously described in terms of "momentary reaction potential" (Hull, 1943), "reflex strength" (Skinner, 1938), and "retrieval strength" (Bjork & Bjork, 1992)[1]. Learning is frequently assessed by collecting MOPs and then analyzing/plotting them across time, across tasks, or both to infer changes in knowledge, understanding, and skills.

During initial skill acquisition, one can only observe changes in short-term performance. Unfortunately, observable performance during training is a poor indicator of long-term durability or flexibility. This is because during initial skill acquisition, the relatively permanent learning effects are obscured by temporary performance-enhancing effects that dissipate over time, such as when the training is withdrawn or when the critical retrieval cues from the learning environment are no longer present in the transfer environment (Schmidt & Bjork, 1992). If one only looks at immediate measures of performance, one would mistakenly favor sub-optimal learning strategies over more optimal ones (Bjork & Bjork, 2011; Christina & Bjork, 1991). Indeed, certain learning strategies known as "desirable difficulties in learning" – which include varied practice conditions during training, interleaved training content, and variable or infrequent feedback – actually decrease short-term performance during skill acquisition, but substantially improve long-term durability and transfer (Bjork & Bjork, 2011; Schmidt & Bjork, 1992). In the following sections, we describe how learning has been variously operationalized in the psychological literature. We then describe how we operationalized MOLs and MOPs in the current study.

[1] We recognize that researchers differ in the definitions of learning and performance. In this paper, we use the definitions proposed by Bjork and colleagues.

MEASURES OF DURABILITY, RETENTION, OR RESISTANCE TO FORGETTING

Mastery Learning (also known as Original Learning). One way to assess durability is to compare performance against a pre-defined criterion of mastery. The criterion could be a "minimal" level of mastery, such as the first trial performed without error. Using this definition, learning might be defined as correctly catching the 3-wire during a carrier landing. Alternatively, the criterion could be a "greater than minimal" level of mastery, which is defined as multiple successive trials without error (Christina & Bjork, 1991). Using this definition, learning might be defined as catching the 3-wire on five consecutive landings. Since the level of mastery is defined by the number of successful criterion trials, achieving mastery at the more stringent level requires additional practice; by extension, it is expected to produce higher levels of learning.

Overlearning. Another way to measure durability is to calculate the amount of overlearning. Overlearning refers to continued task practice after reaching the pre-defined criteria of mastery. Overlearning is generally calculated as the number of successful post-plateau practice trials divided by the number of trials required to achieve mastery. For example, if a pilot required 10 trials to achieve mastery (catching the 3-wire) and then successfully completes another 5 trials, that pilot's level of overlearning would be 50%. The amount of overlearning is usually compared to the rate of reacquisition after a period of nonuse, for example by comparing the reacquisition curves of participants who received overlearning versus matched controls who did not.

Rates of Skill Acquisition, Decay, and Reacquisition. Still another measure of durability is to assess the rates of skill acquisition, decay, and reacquisition. For example, one could train a sample of learners to achieve a pre-defined level of mastery and plot the number of trials until mastery has been achieved. Candidate time periods of non-use would then be identified (e.g., two weeks of non-use, one month of non-use, two months of non-use) based on theory and/or practice. Random subsets of learners would then be re-tested at each time period. The resulting skill decay curves would be used to identify appropriate refresher training intervals, and subsequent re-acquisition curves would be computed to identify the amount of refresher training required.

Automaticity of Performance. Automaticity can be assessed using the dual-task paradigm (Christina & Bjork, 1991). Using this method, the learner performs two tasks simultaneously in order to assess the amount of available working memory. For example, during a series of simulated carrier landings, the pilot could be required to press the trigger in response to a series of specific visual cues. Automaticity is said to be achieved when the learner achieves mastery at performing both tasks simultaneously, such that neither task interferes with the other.

Cognitive Load. Learning can also be assessed using unobtrusive methods. With increased levels of expertise, cognitive processing for certain tasks should become less controlled and more automatic. This should result in increased amounts of available working memory (WM) and corresponding changes in neurophysiological indicators, such as heart (electrocardiogram; ECG), brain (electroencephalogram; EEG), and eye (electrooculogram; EOG) activity (Fairclough, Venables, & Tattersall, 2005).

MEASURES OF TRANSFER OR FLEXIBILITY

Changes in Task and Environmental Conditions. Flexibility is frequently assessed by having the learner perform a variant of the original task, to perform the original task under different environmental conditions, or a combination of both (Christina & Bjork, 1991; Soderstrom & Bjork, 2013). Sometimes, the transfer trials are performed immediately after training (a "near transfer" research design); other times, they are performed after an extended period of time (a "far transfer" research design). For example, after successfully performing a series of practice trials in the simulator, pilot performance could be assessed during FCLP weeks later. Transfer studies are often difficult to conduct because of logistical and cost concerns, which explains why they are rarely documented in the scientific literature. However, it is not necessary that the transfer task recreate the entire trained task. For example, if the goal were to specifically assess the pilots' perceptual skills, a part-task simulation – such as a temporal occlusion task (Ward, Farrow, Harris, Williams, Eccles, & Ericsson, 2008) – might suffice.

Development of Expert-like Strategies. Other measures of flexibility may involve the development of expert-like behaviors or strategies. In Naval aviation, experienced pilots are able to anticipate the likely course of future events (such as the presence of axial wind over the landing area) by reading the wake and sea state as they set up their landing pass. They then proactively position the jet to be in the best possible position for landing. Similarly, expert

pilots become adept at simultaneously controlling glideslope, angle of attack, and lineup during the landing approach, rather than adjusting one at a time. In aviation parlance, these strategies are colloquially referred to as "proactive flying." The development of such expert-like behaviors is clear evidence of learning.

EXPERIMENTAL DESIGN

As noted previously, the purpose of this study was to assess the extent to which the targeted simulator enhancements improved the learning of carrier landing skills. Therefore, we designed an experiment which compared the targeted simulator upgrades (which included improved visual resolution and ocean wake rendering, IFLOLS size and responsiveness, and improved correlation among visual and motion cues) against a baseline simulator setup which closely represents the Tactical Operational Flight Trainer (TOFT). Since the generalizability of the study findings was a concern, we purposely included a mix of novice (less than 75 carrier landings each) and expert (greater than 200 carrier landings each) pilots. We also developed 12 landing passes that present a variety of day/night, visibility, weather, sea state, and wind conditions.

Due to limited resources, however, a far-transfer task – assessing the pilots' skills during FCLP or actual carrier landings – was not feasible. Moreover, due to the pilots' operational and training schedules, it was unlikely that we would have access to the same pilots at a later point in time. We could only request their participation for two consecutive days of data collection. Therefore, we designed a 2 (standard vs. enhanced visuals) x 2 (motion vs. non-motion) x 2 (expert vs. novice) x 3 (nominal vs. moderate vs. high difficult landing pass) experimental design. Visuals were a within-subjects manipulation. Each pilot flew 12 passes per day (6 passes, a 5-minute break, 6 more passes), and the ordering of image generators was counterbalanced across days. Motion was a between-subjects manipulation. If a pilot received motion, he or she flew with motion on both days. Expertise was a between-subjects variable that was not experimentally manipulated; rather, fleet participants were recruited from each of the two designated expertise groups. Finally, scenario difficulty was a within-subjects manipulation. All pilots flew a mix of nominal- (n=4), moderate- (n=4), and high-difficulty (n=4) passes per day; the order of which was randomized. In the following sections, we describe our experimental procedure and measures in greater detail.

PROCEDURE

On day 1, participants arrived at the MFS laboratory for their individually-scheduled sessions. After providing informed consent, the participant completed a brief background questionnaire and a temporal occlusion pre-test. The participant was then led to the simulator area, was outfitted with unobtrusive ECG sensors, and completed the Multi-Attribute Task Battery (MATB) workload calibration task (Santiago-Espada, Myer, Latorella, & Comstock, 2011) using a standard PC laptop.

The participant then entered the F/A-18 C/D simulator, where the research team calibrated the eye-tracking system. He was then given 5 minutes of free flight and 3 practice landings under nominal conditions to become familiar with the simulator. The participant then completed two blocks of 6 landing passes (randomly ordered), with a 5-minute break in between. During each of the twelve landing passes, we unobtrusively collected simulator output, eye-tracking measures, workload measures, and Landing Signals Officer (LSO) ratings. After completing all twelve passes, the participant completed a brief end-of-day questionnaire which inquired about the particular simulator configuration (based on the unique combination of motion and visual cues) that was flown. He was then released for the day. The total time requirement for day 1 was approximately 90 minutes.

On day 2, the participant again arrived for his individually-scheduled session. His baseline level of workload was re-calibrated, as was the eye-tracking model. The participant then completed 2 blocks of 6 landing passes (randomly ordered), again with a 5-minute break in between. Next, the participant completed the temporal occlusion post-test, as well as an end-of-day questionnaire. Finally, the participant was thoroughly debriefed about the purpose of the experiment. The total time requirement for day 2 was approximately 90 minutes.

PARTICIPANTS

Based on the results of a statistical power analysis, we recruited fifteen Navy F/A-18 aviators to participate in the experiment. All participants were male. The sample had a mean of 7.27 years (s.d. = 3.17) active duty service, a mean flight time of 957.47 hours (s.d. = 670.99) in all aircraft, a mean flight time of 241.67 hours (s.d. = 215.74) in

the F/A-18 C/D, and a mean of 48.13 (s.d. = 66.86) carrier landings. Table 1 provides a summary of the participants' background characteristics by expertise group.

Table 1. Sample Characteristics as a Function of Expertise

	Novice (n=8)	Expert (n=7)
Active Duty Military Service	5.25 (1.58) years	9.57 (2.99) years
Total Flight Time	433.75 (223.92) hours	1556.00 (456.49) hours
Flight Time in C/D	156.88 (80.04) hours	338.57 (283.86) hours
Carrier Landings in C/D	16.00 (19.74) landings	84.86 (83.80) landings

Note: Mean values are displayed. Standard deviations appear in parentheses.

It is important to note that the experts' primary aircraft was the E/F, not the C/D in which they were tested. Navy policy dictates that E/F pilots fly a minimum number of hours in the C/D to maintain qualification in that aircraft. As a result, the statistical breakdown depicted in Table 1 understates the experts' true level of expertise, because it does not include their landings in the E/F.

MEASURES

Simulator-based Measures (mastery). Starting approximately .75 nautical miles (NM) behind the ship, pilots fly straight in for landing. This 15-18 second window of time is called "the groove." For each pass, we calculated the mean absolute deviation (MAD) from the ideal angle of attack (AoA), glide slope (GS), and center line (CL) starting at the beginning of the groove and continuing until touchdown. All 3 variables were measured in degrees and sampled at 40 Hz. Since the measures are represented as deviation scores, smaller numbers indicate better performance.

Expert Observer Ratings (mastery). Each pass was rated by a Navy (reserve) LSO, who was providing contract support to the project. Using standard Navy guidance (U.S. Navy LSO School, 2009), the LSO rated the participants' performance using the standard 5-point scale, with larger numbers indicating a safer, more controlled landing.

Physiological Measures (cognitive load). Prior to entering the simulator, all participants completed a series of standardized tasks using the MATB. During this time, their baseline level of workload was unobtrusively measured using the *Brain Products Quick Amp*, which was equipped with three electrodes and custom-developed workload analysis software (Engler, Schnell, & Walwanis, 2013). Their actual level of workload was then recorded unobtrusively for every pass. In addition, the pilots' visual scan patterns were assessed using the *Smart Eye Pro* eye-tracking system, which was equipped with 3 cameras and custom-developed eye-tracking software (Schnell & Engler, 2014).

Transfer Task Performance (transfer). Prior to entering the simulator, all participants completed a temporal occlusion test (TOT; Ward et al., 2008) using a PC laptop that was equipped with 1440 x 900 pixel WXGA graphics accelerator (pre-test). Immediately after completing the last pass on day 2, all participants complete a second TOT (post-test). Each TOT presented the participant with fourteen 8-second video clips, each of which depicts part of a realistic carrier approach or landing. At the end of each clip, the screen was masked and participants were required to indicate (using the mouse) whether the pilot in the video would need to make a standard versus an aggressive correction (Stacy, Beaubien, Wiggins, Walwanis, & Bolton, 2014). Both accuracy and reaction time (RT) were recorded for each clip.

RESULTS

Due to space constraints, this paper only describes the simulator-based MOLs and MOPs. Because the 12 landing passes were randomly ordered per day, it was necessary to aggregate across a subset of passes to ensure that the MOLs were not confounded by the random effects of scenario difficulty. Since the image generator was a within-subjects manipulation, we could only examine intra-day learning. With this in mind, we aggregated the first 6 passes (prior to the 5-minute break) of the day and compared them with the last 6 passes (after the 5-minute break). Since motion was a between-subjects manipulation, we were able to statistically examine inter-day learning. We did this by again aggregating the first and last 6 passes per day, and then plotting performance across both days. By

comparison, performance assessments were made by aggregating across all 24 passes. As will be seen below, the choice of measures – MOLs vs. MOPs – led to very different conclusions.

To assess the effects of learning with regard to visuals, we conducted a multivariate analysis of covariance (MANCOVA) with all three dependent variables (absolute glideslope error, absolute lineup error, and absolute angle of attack error) simultaneously. As noted previously, since the combined dependent variable represents deviations from the ideal flight path and attitude, lower values reflect better performance. After statistically controlling for the effects of scenario difficulty, we found a statistically significant learning x image generator interaction ($F_{(2,316)}$ = 2.09, p. = .05). The results, graphically summarized in Figure 1, suggest that under the typical image generator, experts outperformed novices, and there was no learning for either group. Under the enhanced image generator, experts initially outperformed novices, whose performance was initially degraded. However, by the end of the day, the novices overcame the initial perturbation, and their performance became statistically indistinguishable from that of the experts, as evidenced by the overlapping standard errors. In essence, the only group that demonstrated learning was novices in the enhanced image generator condition. This change in performance over time is .42 standard deviations, which corresponds to a "small" to "medium" learning effect, according to statistical rules of thumb (Cohen & Cohen, 1982).

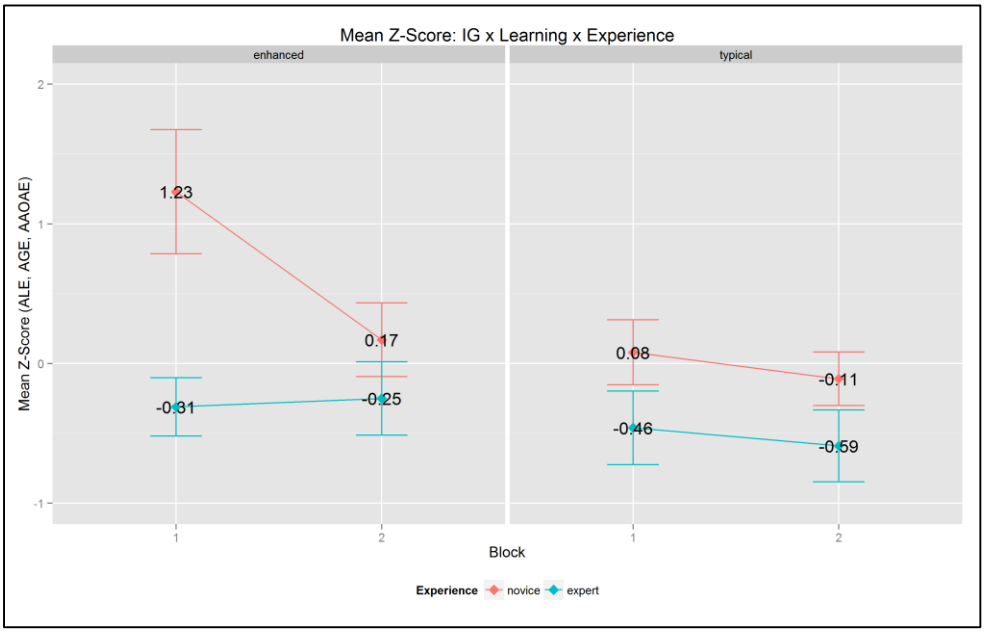

Figure 1. Measures of Learning (Visuals)

To assess the effects of learning with regard to motion, we again conducted a multivariate analysis of covariance (MANCOVA) with all three dependent variables (absolute glideslope error, absolute lineup error, and absolute angle of attack error) simultaneously. After statistically controlling for the effects of scenario difficulty, we found a statistically significant learning x motion x experience interaction ($F_{(4,308)}$ = 4.48, p. < .001). The results, which are graphically summarized in Figure 2, suggest that experts demonstrated no learning, regardless of motion. For novices, however, learning varied as a function of motion. With motion off, novices demonstrated intra-day learning, but the gain dissipated across days. With motion on, the performance of novices was initially degraded. However, by the end of the two-day experiment, the novices overcame this initial perturbation and their performance became statistically indistinguishable from that of the experts, as evidenced by the overlapping standard errors. In essence, the only group that demonstrated learning was novices in the motion condition. This change in performance over time is .89 standard deviations, which corresponds to a "large" learning effect, according to statistical rules of thumb (Cohen & Cohen, 1982). Both results – for the visuals and motion – are consistent with the "desirable difficulties in learning" effect identified by Bjork and colleagues (Bjork & Bjork, 2011; Christina and Bjork, 1991). These results are not unexpected, because the participants were given no explicit instruction in how to use the new visual or motion cues to assist them in landing the jet. Moreover, they were given no feedback about their performance after each pass. They had to figure out how to use these new cues on their own.

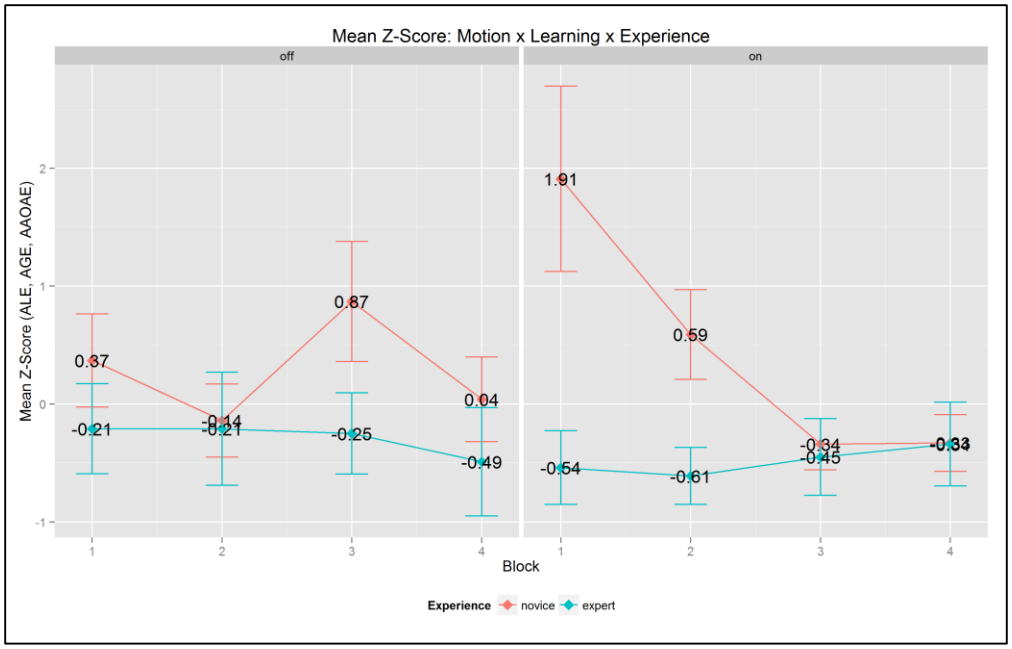

Figure 2. Measures of Learning (Motion)

To assess performance at the end of the experiment, we again conducted a multivariate analysis of covariance (MANCOVA) with all three dependent variables (absolute glideslope error, absolute lineup error, and absolute angle of attack error) simultaneously. After statistically controlling for the effects of scenario difficulty, we found a marginally significant learning x motion x experience interaction ($F_{(1,316)} = 2.62$, p. = .106). The results, which are graphically summarized in Figure 3, suggest that experts outperformed novices in all cases (again, smaller numbers indicate better performance). Moreover, the typical image generator was superior to the enhanced image generator. Finally, the effects of motion were mixed. These last two effects – for visuals and motion – are counter to what we observed for the learning analysis, and caused by aggregating across all 24 passes simultaneously. In essence, by not analyzing performance across trials – to infer if learning has occurred – the results conflate the initial perturbations with the subsequent learning gains, thereby providing an inaccurate representation of what actually occurred.

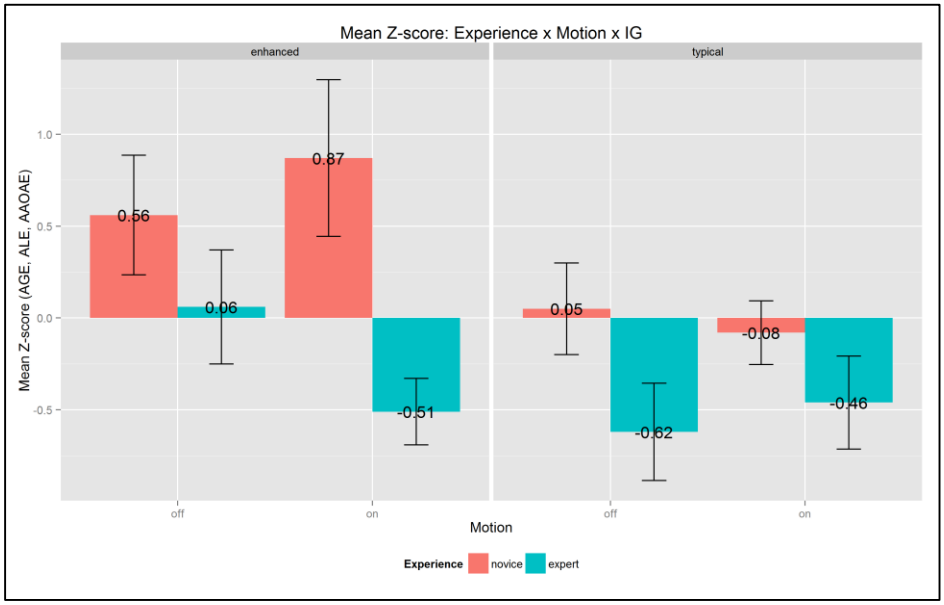

Figure 3. Measures of Performance (Motion and Visuals)

CONCLUSIONS AND IMPLICATIONS FOR PRACTICE

In this study, we sought to assess the effects of targeted simulator fidelity cue improvements on learning carrier landing skills. The choice of fidelity cue upgrades was made based on a combination of prior research findings (Fowlkes et al., 2011) and Cognitive Task Analysis (CTA)-like methods (Wiggins, 2013). Specifically, the CTA-like methods involved having Subject Matter Experts fly a series of simulated landing passes while "talking aloud." During each pass, their performance was audio- and video-taped. After each pass, the SMEs immediately reviewed the recordings of their performance. During this time, they elaborated on their decision processes, particularly with regard to the fidelity cues that they would normally use at the ship, but which were absent in the simulator. The final choice of fidelity upgrades was driven by several factors: cue salience and diagnosticity; anticipated effects on learning; fit with the study's overall goals (visual and motion fidelity); and resource requirements.

The effectiveness of these fidelity cue upgrades was assessed using a rigorous experimental design with random assignment to condition, counterbalancing, and statistical covariation. The results suggest that the targeted motion and visual upgrades initially perturbed the novices' performance. This is not surprising, because the participants were given no explicit instruction in how to use these new cues. Despite the initial perturbation, the novices demonstrated clear evidence of learning (intra-day learning for the enhanced image generator, and inter-day learning for motion) such that by the end of the experiment, their performance was statistically indistinguishable from that of experts. By comparison, the performance analysis suggested a very different picture. Specifically, the performance analysis suggested that the typical image generator was superior to the enhanced image generator, and that the effect of motion was equivocal. This comparison of MOPs and MOLs highlights the critical need for researchers to explicitly consider the way that they operationalize learning. In the following sections, we offer some practical guidelines that readers can use in their own research.

- *Guideline #1: Measure performance separately for each trial*. Although there are many definitions of learning, we have primarily used the one proposed by Bjork and colleagues, which emphasizes relatively permanent changes in knowledge, understanding, skills, and/or the capacity to respond over time (Soderstrom & Bjork, 2013). In order to assess learning (durability), one needs to measure performance for every trial separately, and then analyze it as a function of time.
- *Guideline #2*: *Include multiple measures of learning, whenever possible*. In this study, we collected simulator-based measures, expert observer ratings, and unobtrusive physiological measures during each pass in the simulator. We also assessed pre-post changes in performance on a transfer (temporal occlusion) task outside of the simulator. We did so because each measure has its own unique strengths and weaknesses. For example, expert observer ratings are the easiest to collect. However, they can be quite coarse, thereby making it difficult to identify small differences over time or between groups. By contrast, simulator-based measures are sampled unobtrusively and without error. However, it takes a great deal of computer programming expertise – and Subject Matter Expert (SME) input – to filter, aggregate, transform, and integrate the raw data streams to create meaningful measures.
- *Guideline #3*: *Statistically control for unwanted variance across trials*. There are two schools of thought when it comes to assessing durability. One is to use the same trial conditions each time. So, for example, we could have had the pilots fly the exact same pass 6 times in a row. The downside, obviously, is that this method limits the generalizability of the results. The other school of thought is that performance should be assessed across different conditions. We chose this option to ensure that the results generalize across a range of landing pass conditions, including weather, visibility, sea state, time of day, and the like. To design our landing passes, we began by identifying all the relevant factors that we wanted to manipulate (e.g., wind speed over the deck), and the different levels of each (e.g., 22 knots, 30 knots, 40 knots). With the example of wind speed, 22 knots was identified by our SMEs as the ideal value. Using that as a baseline, our SMEs identified the other levels (30 knots, 40 knots) as 30% and 60% more difficult vis-à-vis baseline. We did this separately for each variable that we wanted to manipulate (wind speed, visibility, wind shear at pattern altitude, etc.) and summed the percentage values to calculate a numerical "difficulty" score for each landing pass. These difficulty scores were then used as the covariate in our statistical analyses.
- *Guideline #4*: *Specify your statistical model*. In order to assess the learning-related benefits of the upgraded visuals (for example), we had to measure several variables: scenario difficulty (covariate), image generator condition, time, and expertise. In most statistical packages, to assess the interaction of image generator x time x expertise, one would first need to include all main effects and two-way interactions before testing

the three-way interaction, which was of primary interest to our study. However, doing so uses up critical degrees of freedom. In performing our analyses, we specified our statistical model to ignore the main effects. Certain statistical packages, such as R (cran.r-project.org) allow the researcher a great deal of flexibility in specifying the model that is to be tested. We encourage researchers to do so, rather than being constrained by other packages which do not permit such flexibility.

- *Guideline #5: Use experts as a comparison group for quantifying the extent of learning and to rule out "learning the simulator*." If we had only included novices in the study, we could demonstrate that they had learned, but it would be impossible to know if they had learned carrier landing skills, or if they had simply learned "how to fly the simulator." By including a sample of experts, we were able to disentangle these two issues. Specifically, we observed that experts' performance did not change over time; they demonstrated no evidence of learning. This is not surprising because their expertise was gained outside of the simulator. There should be little or no learning with an additional 3 hour of practice. If they had improved over time, it would have been clear evidence that at least one group – the experts – had learned to fly the simulator. While the fact that experts did not learn to "game the simulator" is not conclusive evidence that novices also did not learn to "game the simulator," it does provides indirect evidence that they learned the actual skill in question, not learning the simulator.
- *Guideline #6: Use a part-task simulator as the transfer task, if appropriate.* For the purposes of this study, we were interested in training and assessing perceptual-motor skills. As the name suggests, perceptual-motor skills involve learning patterns of cues in the environment, and recognizing patterns from incomplete information (Stacy, Beaubien, Wiggins, Walwanis, & Bolton, 2014). Therefore, we chose to use the temporal occlusion paradigm as our transfer task. Using actual video clips of simulated flights, we were able to quickly and easily create a test of perceptual motor skills using the freely-available OpenSesame software tool (osdoc.cogsci.nl). Researchers may need to be creative when selecting a transfer task, but it can be done.

Even when faced with rigid resource constraints, such as the ones that we faced in this study – for example, we could only have access to 15 pilots over two consecutive days – it is still possible to disentangle MOLs from MOPs. One just needs a flexible and well-designed data collection approach. The guidelines listed here can help other researchers to apply these methods in their own research, with the ultimate goal of better differentiating between learning and performance.

ACKNOWLEDGEMENTS

This study could not have been completed without the tireless efforts of numerous researchers, simulation engineers, software developers, and F/A-18 SMEs. We extend our sincerest thanks to (in alphabetical order by last name): Laura Biggerstaff, Christopher Cheok, Brett Chladny, Beth "Gabby" Creighton, Robert "Crater" Creighton, Lisa Dang, Timothy Scott Davis, Joseph Engler, Mark "Jitters" Kircher, Steve Moss, Stephen Naylor, Jennifer Pagan, Christopher Parish, David Perdue, Bruce Riner, Thomas Schnell, Sandra Velez, Michael "Sting" Wallace, Kelli Wiuff, and Michael Yocius.

Funding for this study was provided by the Office of Naval Research to Aptima, Inc., NAWCTSD, and NAWCAD. The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of their sponsors at 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

Bjork, E., & Bjork, R. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. Gernsbacher, R. Pew, L. Hough, & J. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56-64). New York: Worth.

Bjork, R. & Bjork, E. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A Healy, S. Kosslyn, & R. Schiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35-67). Mahwah, NJ: Erlbaum.

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.

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.

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.

Fuller, S. (2011). Adapting air operations to energy challenges. In L. Simmons (Ed.), *Adapting to climate and energy challenges: Options for U.S. maritime forces* (pp. 225-232). Laurel, MD: The Johns Hopkins University Applied Physics Laboratory.

Government Accounting Office. (2012). *Navy training: Observations of the Navy's use of live and simulated training*. Report No. GAO-12-725R. Available: http://www.gao.gov/assets/600/592056.pdf.

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

Santiago-Espada, Y., Myer, R., Latorella, K., & Comstock, J. (2011). *The multi-attribute task battery II (MATB-II) software for human performance and workload research: A user's guide*. Hampton, VA: NASA Langley Research Center.

Schmidt, R., & Bjork, R. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207-217.

Skinner, B. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.

Soderstrom, N., & Bjork, R. (2013). Learning vs. performance. In D. Dunn (Ed.), *Oxford Bibliographies Online*. doi:10.1093/OBO/9780199828340-0081

Stacy, E., Beaubien, J., Wiggins, S., Walwanis, M., & Bolton, A. (under review). *Using temporal occlusion to assess carrier landing skills*. Submitted for presentation at the 2014 Interservice/Industry Training, Simulation, and Education Conference.

U.S. Navy LSO School. (2009). *Grading concepts and debriefing techniques*. Virginia Beach, VA: Naval Air Station Oceana.

Ward, P., Farrow, D., Harris, K., Williams, A., Eccles, D., & Ericsson, A. (2008). Training perceptual-cognitive skills: Can sports psychology research inform military decision training? *Military Psychology, 20 (Supplement 1)*, S71-S102.

Wiggins, S. M. (2013). *Carrier cue qualification workshop*. Unpublished technical report. Woburn, MA: Aptima, Inc.