2015

Evaluation of an Eye Tracking-Based Assessment and Debrief Tool for Training Next Generation Multirole Tactical Aviation Skills

Meredith Carroll

Glenn Surpris

Greg Sidor

Winston Bennett Jr.

EVALUATION OF AN EYE TRACKING-BASED ASSESSMENT AND DEBRIEF TOOL FOR TRAINING NEXT GENERATION MULTIROLE TACTICAL AVIATION SKILLS

Meredith Carroll & Glenn Surpris, Design Interactive, Inc., Orlando, FL
Greg Sidor & Winston Bennett, Jr., Air Force Research Lab, Dayton, OH

As tactical aircraft become increasingly complex, pilots' cognitive resources will become increasingly strained, especially as more critical and multifaceted information is presented on Helmet Mounted Displays (HMDs). Therefore, it is critical to ensure training results in pilots learning optimal strategies for operating in this information-rich environment, including appropriate attention allocation between different dynamic, adjustable displays and efficient scan strategies. To achieve this, a performance assessment and debrief system was developed that incorporates eye tracking technology into an HMD-enabled multirole fighter simulation to capture and process gaze data to aid in diagnosing why a pilot error occurred. The system utilizes eye tracking to 1) measure attentional focus and scan patterns, 2) diagnose errors in performance and attention allocation, and 3) display performance and attention allocation summaries and mission replay overlaid with pilot scan patterns. A training effectiveness evaluation of the system was conducted with 14 Air National Guard F-16 pilots at the 180th Fighter Wing in Toledo, OH. The F-16 is a current generation multirole aircraft with complicated displays and a variety of mission data displays available for mission execution. Participants were split into two conditions, including a control group which received debriefings utilizing traditional mission replay tools and an experimental group which received debriefings which utilized pilot scan data presented in conjunction with traditional mission replay tools. Results suggest that debriefs utilizing pilot scan data have the ability to support pilots in more quickly adjusting their scan strategies to those most optimal for performance in future, more data intensive tactical fighter environments. The paper will present the system design, experimental methods and a discussion of the results and implications for training.

According to the Air Force Transformation 2010, "…the ultimate source of air and space combat capability resides in the men and women of the Air Force. …[Our] first priority is ensuring they receive the precise education, training, and professional development necessary to provide a quality edge second to none" (United States Air Force, 2010, p. 6). Within the Department of Defense, the ever increasing amount and complexity of knowledge, skills and abilities (KSAs) demanded of their personnel has created the need to develop and utilize tools to increase the efficiency and effectiveness by which training takes place. This is especially true for next generation tactical aircraft. Next generation multirole fighter aircraft will increasingly use a Helmet Mounted Display (HMD) as a primary instrument and sensor display. This comes in conjunction with a significant increase in duties required by the pilots and has the potential to put incredible strain on the pilot's cognitive resources by exposing him to large amounts of data from disparate sources that can quickly exceed natural cognitive processing limits. It is critical to ensure training results in pilots learning optimal strategies for operating in this information-rich environment, including appropriate attention allocation between the different displays and effective and efficient scan strategies. Key to this may be the integration of pilot scan data into assessment and debrief. Visual attention can provide important insights to the information used in task performance, such as the importance of various features or cues (Raab & Johnson, 2007). Several studies (Raab & Johson, 2007; Jarodzka, Scheiter, Gerjets, & van Gog, 2009; Mello-Thoms et al., 2008; and White, Hutson, & Hutchinson, 1997) have demonstrated that eye tracking can aid in the assessment of perception through measurement of visual attention during observation via gaze, scan path, and fixation data. These measures can provide a means for increasing the granularity of performance feedback and a means by which pilots can understand and adjust their scan strategies. Findings in the literature indicate experts have very well-defined scan patterns (Burgert et al., 2007; Jarodzka et al., 2009; and Kasarskis, Stehwien, Hickox, Artez, & Wickens, 2007). This is the result of experts having developed very strong systematic scan strategies in comparison to novices who have less structured scan patterns. Novice scan patterns are typically influenced most by bottom up processes which draw attention to salient features in the environment (Jarodzka et al., 2009). It was hypothesized that, given their experience, it may prove a challenge for legacy pilots such as F-16 pilots to quickly transition their scan strategies to those most optimal for next generation multirole fighter operations. While next generation fighters have similar tactical missions, there is a great deal more and different information presented on the HMD to support those same tactical operations. These pilots will need to redevelop scan patterns to those more optimal for operations and systems of the new multirole fighter aircraft. Thus it was hypothesized that pilots who

receive training debriefs that incorporate feedback related to scan patterns and attention allocation will more quickly alter their scan strategies to those most optimal for the new aircraft operations and related information displays, compared to pilots who receive more traditional debrief methods based on performance measures and serial mission replays. It was also hypothesized that these changes would lead to greater performance improvements in pilots who received eye tracking-based debrief. An experiment was conducted to test these hypotheses.

Method

Participants

Reserve and active duty Air Force pilots (14 men, M_{age} = 36.5 years, *SD* = 5.8) stationed at the Ohio Air National Guard 180th Fighter Wing in Toledo were recruited by email for voluntary participation in the research. Participants were not compensated in addition to their regular salaries; rather, their participation was scheduled during their normal duty day. All participants were required to have either normal or corrected to normal vision by use of contact lenses only, not glasses. Pilot rank ranged from First Lieutenant to Colonel, with an average of 13.5 years (*SD* = 7.3) of military service and 11 years (*SD* = 6.2) of F-16 experience. Participants completed an average of 1800 flight hours (*SD* = 965.2) in the F-16. Two participants had next generation fighter simulator experience, one with 4 hours of experience and another with 175 hours.

Experimental Design and Measures

The experiment used a between subjects repeated measures design. All participants received a pretest in which they performed a series of tactical engagements followed by a debrief per their condition (eye tracking-based vs. traditional). Then each participant performed two brief tactical training scenarios followed by a debrief per their condition. All participants then received a posttest in which they performed a series of tactical engagements. Dependent Variables measured include Instructor Evaluation of Trainee Performance (Correct ID, Shots Valid, Hits, Survivability, Flow Errors, Switch Errors, Communication Errors, Display Utilization, Target Detection, ID, Employment, Cold Ops, Merge Prep, Battle Damage Assessment, Overall Student Performance), System collected measures (Missile Result and Shooter Loft Angle) and attention allocation/scan strategy measures (# fixation and average fixation duration on high/low priority areas, time spent heads up vs. heads down).

Testbed

The testbed utilized Helmet-Mounted Display ASsessment System for the Evaluation of eSsential Skills (HMD ASSESS), a performance assessment and debrief system that incorporates eye tracking technology into an HMD-enabled next generation fighter simulation to capture and process scan data to aid in diagnosing why a pilot error occurred. HMD ASSESS was integrated with a desktop simulation that allows students to interface with the displays and controls via a flight representative hands on throttle and stick (HOTAS; See Figure 1).

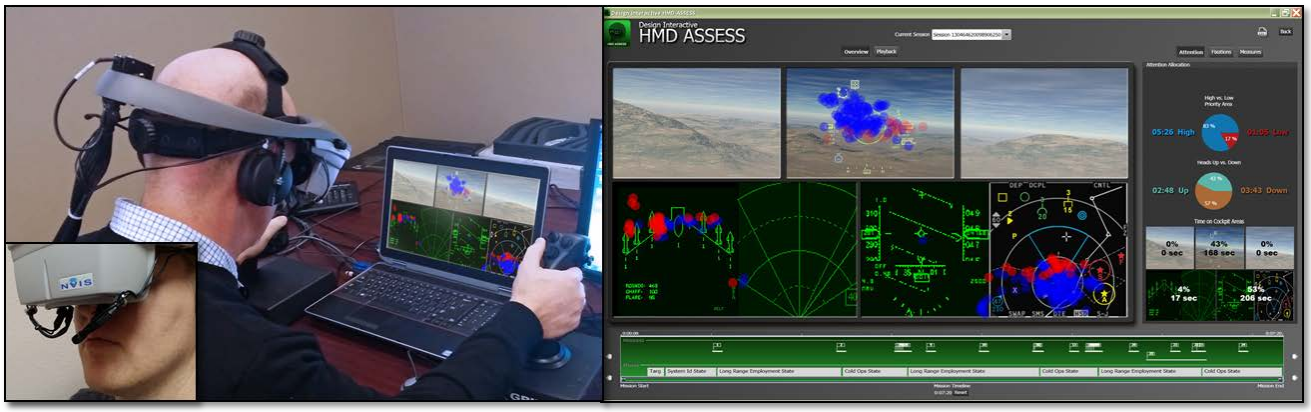

Figure 1. HMD ASSESS Prototype (left) and Instructor Displays (right) with notional iconography and data for the purposes of the paper.

As a pilot flies within the HMD-enabled simulator, the HMD ASSESS measurement component captures gaze tracking measures utilizing the Viewpoint EyeTracker® from Arrington Research Inc. The ViewPoint EyeTracker® integrates miniature high resolution cameras with small infrared lights which allow the determination of pupil location based on corneal reflections. As a pilot monitors different displays within the simulator, the eye

tracker determines the X, Y coordinate associated with each gaze point which is mapped to an area of interest (AOI) (i.e., instrument, display) using a pursuit algorithm which tracks the AOIs as they move across the screen. Behavioral measures from the simulation are also captured, including: 1) simulation events (e.g., flight control inputs, missiles fired, Integrated Caution and Warning System (ICAWS) messages), 2) flight/weapons parameters (e.g., altitude, airspeed, heading, digital maneuvering cue (DMC)), and 3) entity states (e.g., bandits alive, ownship alive). The effectiveness of a pilot's scan and visual attention allocation strategies is then diagnosed by determining where a pilot's attention is fixated and if these fixations intersect with high priority instruments and displays for the task he is currently performing. The high priority displays and instruments for the different segments or phases of the tactical scenarios were determined and defined in a state-engine a priori by fighter pilot subject matter expertise. Additionally, the system determines whether each fixation is associated with the pilot being heads down (i.e., fixating on the panoramic cockpit displays (PCD) panels) or heads up (i.e., fixating out the window). The results are presented in both an interactive multi-level mission timeline overlaid with pilot performance and scan data summaries and an audio/video mission replay with scan data overlays to illustrate the context surrounding errors.

Procedure

All participants were run over the course of six consecutive days at the Ohio Air National Guard facilities. Each experimental session lasted between two and three hours. After reading and signing the informed consent, all participants received familiarization training on the desktop simulator. The familiarization training consisted of a Power Point presentation given by a SME that detailed functionality of the displays and controls necessary to complete the target scenarios. Each participant then donned the HMD and a fitting procedure was performed lasting approximately 5 to 10 minutes. Each participant then completed one familiarization scenario (also approximately 5-10 min) in which they practiced flying the aircraft and targeting enemies. After the familiarization scenario, the eye tracker was adjusted and calibrated for the participant. This procedure involved adjusting the eye-tracking cameras and lighting to get a fix on the participant's pupils. After calibration, participants performed a series of four scenarios containing either two or three adversaries. Each scenario was an air-to-air combat scenario in which the participant had to identify, fix, track and engage enemy aircraft. During each scenario, the instructor played the role of Airborne Warning and Control System (AWACS), providing bullseye picture calls to the trainees. Prior to each scenario, the eye tracker was re-calibrated for those in the experimental condition.

For the control group, the eye tracker was only calibrated for the pretest and posttest scenarios. Eye tracking data was collected for all participants during these scenarios but was not presented to participants in the control condition. Following each scenario, the participant removed the HMD and was debriefed by an instructor on his performance facilitated by the HMD ASSESS debrief system. Participants in the control group received debriefs from the SME who utilized the HMD ASSESS display with eye tracking data disabled. The instructor utilized the HMD ASSESS playback mode to playback the mission, utilizing both the video replay and the values present in the parameters tab to provide feedback. Specifically, participants were provided feedback on the validity (ID and DMC value) of each of their missile deployments as well as the result of their missile deployments. Additionally, participants received feedback on how many red shots were directed towards them; how many flow, switch, and communication errors they made; and tips for avoiding these errors. If the participant did well in any of these categories, they also received positive feedback. For any errors identified, the instructor provided recommendations for improving performance related to these errors in future scenarios. For example, in the control debrief, if a participant lost too much altitude during an out maneuver, the instructor told the participant to note his rate of descent and altitude reading during that phase. Participants in the HMD ASSESS group received a debrief in the exact same format; however, the eye tracking data was enabled. This allowed the instructor to also provide definitive feedback on the root cause of many of the errors. For example, if a participant lost too much altitude during an out maneuver, the instructor was able to see where the participant's visual attention was focused (e.g. tactical situation display), provide feedback on why it was no longer an high priority area to monitor, and instruct them to be mindful of their rate of descent and altitude read outs on future missions. The instructor also used the eye tracking data to validate items he was unable to with the control group. For example, the instructor was able to verify with eye tracking if a participant viewed the Expanded Data Window while making target identifications. The participant also received positive feedback for vigilant scan patterns that incorporated all of the necessary AOIs for a given task. Debriefs were scripted and standardized across both conditions as much as possible. The participant was then fully debriefed with information about the study, including any possible effects the experiment will have and information for access to the results.

Results

Preliminary analyses were conducted to identify potentially confounding variables. One participant was omitted for having 175 hours in a next generation multirole fighter simulation. A multivariate analysis of variance (MANOVA) was performed with a between groups factor of treatment condition (eye tracking-based vs. traditional debrief) for dependent variables (DV) of age, rank, service years, F-16 experience in years and flight hours, F16 training hours – live and simulation, and next generation multirole fighter training hours – live and simulation. There were no significant differences between the two groups in the above demographic variables. A multivariate ANOVA was performed with a between groups IV of treatment condition (eye tracking-based vs. traditional debrief) for DVs of durations associated with each of the four scenarios and each of the four debriefs. There were no significant differences indicating participants in the two training groups received training over approximately the same amount of time.

Scan Data: Next, eye tracking data was analyzed to determine if debrief with scan data led to improved pretest to posttest changes in a trainee's scan strategies compared to the control group. Utilizing 12 of the 14 participants (7 experimental, 5 control; the participant with 175 hours of next generation multirole fighter simulation training was omitted and pretest eye tracking data was lost for one participant), a repeated measures MANOVA was performed with a within groups factor of trial (pretest vs. posttest) and a between groups factor of treatment condition (eye tracking-based vs. traditional debrief). To account for differences resulting from eye tracking data quality and not the treatment condition, calibration quality and eye tracking quality scores were utilized as covariates. As predicted, when comparing pretest to posttest scan data, participants receiving eye tracking-based debriefs altered their scan patterns to focus more on high priority areas as opposed to the control group. These differences were not statistically significant, but trended towards significance with moderately high effect sizes. Specifically, there was an interaction between trial and condition for time spent fixating on high priority areas (F (1, 6) = 3.03, p = .13, η^2= .34), time spent fixating on low priority areas (F (1, 6) = 3.93, p = .09, η^2= .39), number of fixations on high priority areas (F (1, 6) = 2.07, p = .20, η^2= .26),), and number of fixations on low priority areas (F (1, 6) = 1.59, p = .25, η^2= .21). From pretest to posttest, participants receiving eye tracking-based debriefs increased the time spent fixating and their number of fixations on high priority areas while decreasing the time spent fixating and their number of fixations on low priority areas. The control group displayed inverse patterns decreasing time and fixations on high priority areas while increasing time and fixations on low priority areas (see Figure 2). Additionally, there was an interaction between trial and condition that trended towards significance for time spent fixating heads down (F (1, 6) = 3.74, p = .10, η^2= .38), with participants receiving eye tracking-based debriefs decreasing the time they were heads down from pretest to posttest and control participants increasing the time they spent heads down from pretest to posttest.

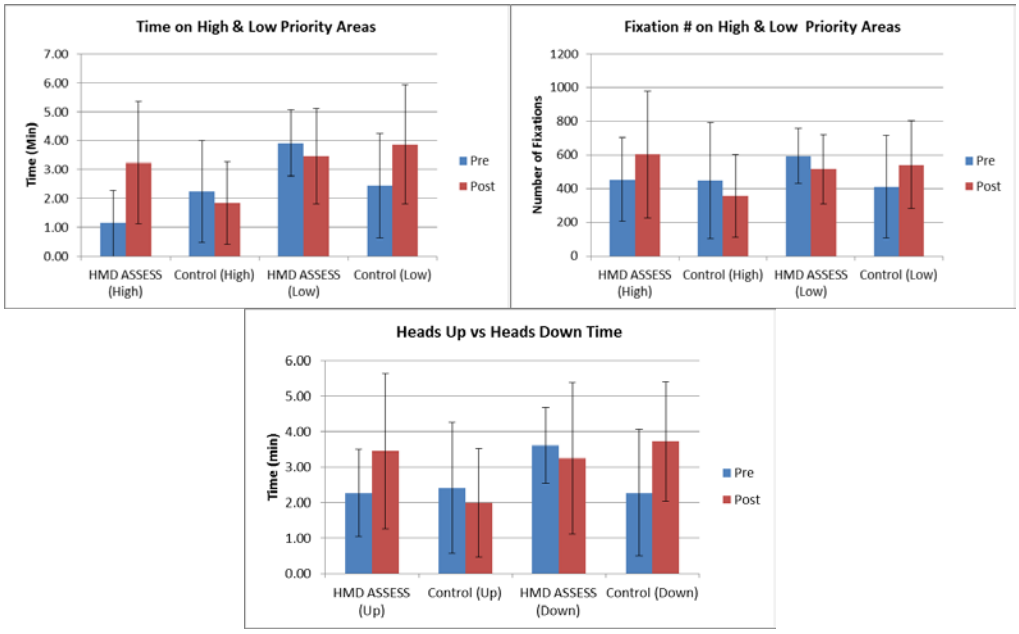

Figure 1. Scan data results.

Performance: These differences in scan strategies did not translate to performance differences. Instructor evaluation measures were analyzed to determine if debriefs with scan data led to improved pretest to posttest changes in a pilot's performance compared to the control group. Utilizing 13 of the 14 participants (8 experimental, 5 control; the participant with 175 hours of next generation fighter simulation training was omitted), a repeated measures MANOVA was performed with a within groups factor of trial (pretest vs. posttest) and a between groups factor of treatment condition (eye tracking-based vs. traditional debrief). There was a significant trial effect for multiple measures with both groups showing improvement from pretest to posttest in overall mission performance (F (1, 11) = 6.04, p = .02, η^2= .39), number of valid shots (F (1, 11) = 12.49, p = .01, η^2= .53), location and utilization of controls and displays (F (1, 11) = 24.5, p = .00, η^2= .69), and hits (F (1, 11) = 9.59, p = .01, η^2= .47 (see Figure 3). There were no between group differences or interactions, suggesting that there were not differential training performance effects between the groups. No other instructor-based measures showed significant differences.

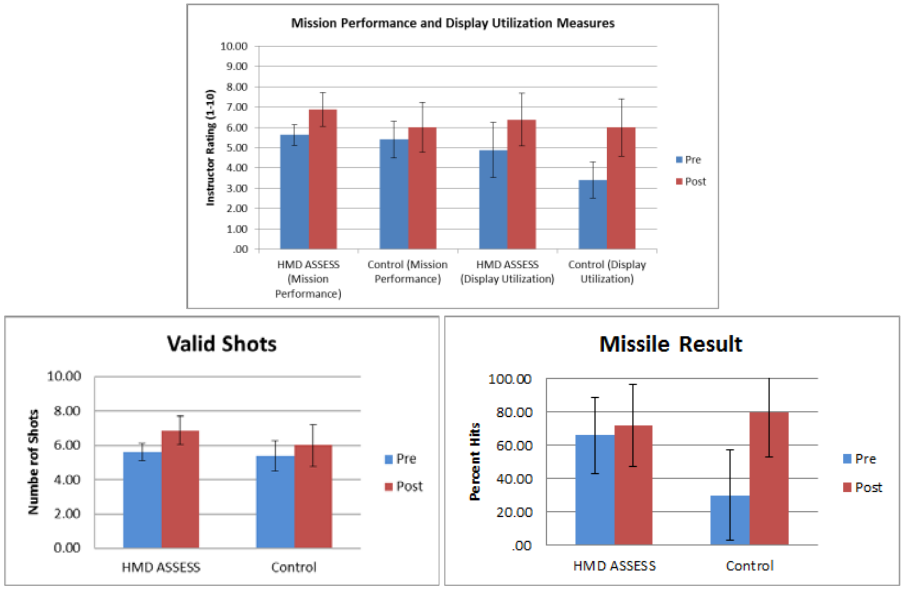

Figure 3. Instructor evaluation measures and system collected measures.

Similar results were seen in system collected performance measures. Due to limitations in the amount and accuracy of data being published by the desktop simulator on the Distributed Interactive Simulation network, only two measures could be calculated across all missiles fired in the pretest and posttest scenarios: Missile Result (0 = miss, 1 = hit) and Shooter Loft Angle (in degrees). These measures were analyzed to determine if debrief with scan data led to improved pretest to posttest changes in a trainee's performance compared to the control group. Utilizing 13 of the 14 participants (8 experimental, 5 control; the participant with 175 hours of next generation fighter simulation training was omitted), a repeated measures multivariate ANOVA was performed with a within groups IV of trial (pretest vs. posttest) and a between groups IV of treatment condition (eye tracking-based vs. traditional debrief). There was a significant trial effect for Result (F (1, 11) = 12.69, p = .01, η^2= .54), with both groups improving from pretest to posttest. There was also a significant interaction for Result (F (1, 11) = 7.68, p = .02, η^2= .41), with control group participants showing greater increases from pretest to posttest, although the control group had a significantly lower pretest score, with posttests being only slightly higher than the experimental group (see Figure 3). There were no between group differences. There were also no significant within or between group differences with respect to Shooter Loft Angle.

Discussion

These results provide promising insight into the potential for eye tracking data to improve the training effectiveness of current tactical aviation debriefs. The results suggest that debrief utilizing scan data has the ability to support pilots in more quickly adjusting their scan strategies to those most optimal for performance. Pilots receiving eye tracking-based debriefs increased the time and number of fixations on high priority areas and the time spent heads up, patterns that were not seen in the control group. Although these did not translate into greater performance improvements over traditional debrief methods, this may be due to the very limited amount of training received or the sensitivity of the performance measures collected. Pilots were trained on a total of four brief

scenarios including the pretest and posttest. This resulted in approximately an hour of training time including debrief. This may not have been enough time for these changes in scan strategies to translate to actual performance improvements. Results of this study must be considered cautiously due to several limitations of the study. There were a very limited number of participants in the study due to the availability of the pilots to support and the time it took to complete the study. The study was also not a blind study, as the instructor had to know which condition in order to give the appropriate debrief. Further, not only was there an unequal number of participants, but treatment group assignments were not random. Those participants in the beginning who had poor eye tracking calibration data were placed in the control group as debrief with scan data was not possible. As the week progressed and methods for fitting the HMD and adjusting the eye tracking cameras resulted in greater proportions of participants with eye tracking data, participants were eventually split up between the groups randomly. Finally, the adversary behaviors in the pretest and posttest scenarios were not consistent. Despite designing the scenarios such that the enemy behavior was scripted to react a particular way, this was not always the case and was not something the system operators could control.

Conclusion

Eye tracking technology provides the ability to go beyond measuring observable performance outcomes (e.g., did the trainees effectively engage the target?), making feasible the measurement of unobservable perceptual and cognitive processes such as pilot scan and attention allocation. These measures aid in the identification of where in the piloting process breakdowns occur (e.g., pilot failed to monitor certain shot parameters in order to identify optimal time to shoot). This facilitates more tailored feedback, potentially leading to significant improvements in training effectiveness and efficiency. Further research is needed in this area, however, this study provides promising data in support of utilizing eye-tracking assessment and debriefs to train next generation multirole tactical aviation skills.

Acknowledgements

This work was funded by the Air Force Research Lab (AFRL) under contract # FA8650-12-C-6303. The views and conclusions contained in this presentation are that of the authors and should not be interpreted as representing the official viewpoints of AFRL or the US government.

References

Burgert, O., Örn, V., Velichkovsky, B. M., Gessat, M., Joos, M., Strauß, G., ... & Hertel, I. (2007, March). Evaluation of perception performance in neck dissection planning using eye tracking and attention landscapes. In *Medical imaging* (pp. 65150B-65150B). International Society for Optics and Photonics.

Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*(2), 146-154.

Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., & Wickens, C. (2001, March). Comparison of expert and novice scan behaviors during VFR flight. In *Proceedings of the 11th International Symposium on Aviation Psychology* (pp. 1-6).

Mello-Thoms, C., Ganott, M., Sumkin, J., Hakim, C., Britton, C., Wallace, L., & Hardesty, L. (2008). Different Search Patterns and Similar Decision Outcomes: How Can Experts Agree in the Decisions They Make When Reading Digital Mammograms? In *Digital mammography* (pp. 212-219). Springer Berlin Heidelberg.

Raab, M., & Johnson, J. G. (2007). Expertise-based differences in search and option-generation strategies. *Journal of Experimental Psychology: Applied*, *13*(3), 158.

United States Air Force. (2010). *The Edge Air Force Transformation 2010*. Retrieved from http://permanent.access.gpo.gov/lps40477/edgeweb.pdf

White Jr, K. P., Hutson, T. L., & Hutchinson, T. E. (1997). Modeling human eye behavior during mammographic scanning: Preliminary results. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *27*(4), 494-505.