

Assessing the Effectiveness of Virtual Reality in the Training of Army Aviators

Ylli Dalladaku, Jacob Kelley, Brycen Lacey, James Mitchiner, Braden Welsh, and Matthew Beigh

Department of Systems Engineering
United States Military Academy, West Point, NY

Corresponding Author: ylli.dalladaku.ko@westpoint.edu

Author Note: Cadets Dalladaku, Kelley, Lacey, Mitchiner, and Welsh are seniors attending the United States Military Academy, majoring in Systems Engineering and Engineering Management. This report is for a senior capstone design project in the Department of Systems Engineering supporting the United States Army Aviation Center of Excellence's Directorate of Simulation (USAACE DOS) under the advising of Major Matthew Beigh.

Abstract: In an effort to modernize flight training the Army has introduced Virtual Reality (VR) through the Aviator Training Next (ATN) program. This program uses VR to supplement "traditional hands-on" training during the initial stages of Initial Entry Rotary Wing Training, or flight school. ATN's primary goal is to produce aviators of the same quality as those trained via traditional means without adding time to Primary phase. As of April 2020, the program is still experimental with only limited students enrolled. This project's comprehensive statistical analysis suggests that thus far the ATN program produces aviators of the same competence level as traditional flight training; however, future class data is still needed to confirm findings and better refine the curriculum's ratio of live flight to VR hours. USAACE plans to continue the experimental program through the Summer of 2020, when there will be a final decision regarding a more robust implementation.

Keywords: Army Aviation, Virtual Reality, Aviator Training Next, Fidelity

1. Virtual Reality Program Overview and Purpose

In an effort to expedite the training process in early stages of flight school and allow students to gain more flight time and experience with their advanced airframe, the United States Army Aviation Center of Excellence (USAACE) at Fort Rucker, Alabama, introduced VR (Virtual Reality) into flight school through the ATN program. The purpose of the ATN program is to pilot VR in the Army and test its effectiveness and viability for training soldiers, pilots, and other Army personnel on their assigned tasks. The introduction of VR seeks to enhance the training experience and quality of Army personnel to accomplish complex tasks, such as operating a rotary wing aircraft. Through the utilization of low-cost headsets and physical flight controls, the VR flight instruction program seeks to augment some live flight hours spent in the physical aircraft, which could allow for the reprioritization of scarce resources to other aspects of flight training, increasing the proficiency of the average graduate. Flight school begins with initial entry rotary-wing training, where all students learn in the same basic training aircraft, the LUH-72 Lakota. In the subsequent phase students train on their specialized combat aircraft, such as the AH-64 Apache, CH-47 Chinook, or UH-60 Blackhawk. The primary phase is further broken up into two assessments, or "checkrides," each occurring after approximately four weeks of training. Referred to as the "P1" and "P2" checkrides, the assessments consist of a combination of an oral and practical flight examination by an impartial instructor pilot and with grading based on a set rubric and well-defined standards. Currently, the ATN program is only designed to train students during the two primary phases, however, there exists the potential for expansion to other training phases if the technology proves beneficial to student learning.

The ATN program has multiple goals designed to train effective aviators and improve the efficiency of flight school. The primary goal of the program is to produce aviators using VR that are as effective as those produced using traditional flight school instruction. Therefore, the goal of this project is to support or deny the ability of VR to meet this goal. Another goal is to optimize the ratio of VR hours to the live flight hours in a program of instruction that maximizes training effectiveness. Lastly, an ancillary goal of the program is to improve the throughput of aviators in training with this additional resource, the result of which could help minimize the wait times (or "bubbles") between phases that currently plagues the system.

The program was implemented on a selective basis starting in September 2019, with only a subset of ten to fifteen students in each class randomly selected for participation. Approximately two-thirds of the trainees at Fort Rucker are still utilizing the traditional training model. The two limitations to the quantity of students in the program are the number of ATN machines and the number of instructor pilots proficient with the system. Each class is currently operating under a slightly

different ratio of live to VR flight hours as part of a larger research effort to determine the optimal combination in future programs of instruction.

As of April 2020, eight classes have progressed through the P1 phase of instruction. The plan is to continue trialing the program with select classes to collect additional data and refine the ratio of live flight hours to VR hours. Among the USAACE senior leadership the program is currently regarded very positively with the potential to modernize and expedite training at flight school. However, issues with the VR model being incomplete and containing multiple technical flaws, as well as lack of communication between stakeholders, have slowed some of the progress throughout the research period.

2. Background Information on Virtual Reality

Since the early 2000s, VR has been used to aid in the training of employees and students as a means to enhance the quality of the training experience. There are a few relevant topics to explore regarding the ability of VR to effectively train workers on their task. The learning process, the consequences and side effects of VR, and the historical effectiveness of VR must all be understood before the impact of VR in a military setting can be studied.

2.1 Learning with Virtual Reality – Fidelity and Training Transfer

Neurons are cells that receive information from other nerve cells or sensory organs and projects that information to other cells and organs in the body. Information passes from neuron to neuron through an information highway called the synapses. Neurons take the information it receives from the synapses and this determines its output. These connections allow for humans to learn information and part of this is guided by experience. The first way that synapses are formed is when these synapses are over produced, then over time are selectively lost. This is a mechanism that the brain uses to incorporate information from experience. The second way that new synapses form is through synapses addition. This process is driven by experience, which is what VR capitalizes on, and lies at the base of most forms of memory in humans (Committee on Developments, 2004). VR can help people learn because it immerses the user in a virtual world experience, and synapses connections are formed from experience.

Training transfer rate and fidelity are decisive parameters to evaluate the VR technology and its performance. Training transfer refers to the level of adaptation acquired in one domain and its application on other domains. While fidelity refers to the ability of a simulator to reproduce real conditions accurately, high fidelity does not always translate into better training transfer rate. The appropriate level of fidelity is a function of the training objectives, individual tasks and learning level of the trainees. A study found some variables that contributed to the conflicting results: subjects who have different motor skills and cognitive capabilities, and instructors, who play an important role because of evaluation, biases, attitudes, and motivation (Myers et al., 2018). The role of instructors cannot be neglected, as they can effectively increase the training transfer rate. Studies suggest that the role of instructors is important to a point that can be compared to the simulator's impact too (Myers et al, 2018). Ultimately, having the appropriate fidelity level to successfully complete performance tasks will increase the effectiveness of ATN to produce competent pilots.

2.2 Historical Effectiveness of Virtual Reality

Analyzing the use of VR in both the civilian and military sector, the VR technology usage has seen a limited, but successful, implementation as a training tool. One study seeking to test the ability of VR to train physical assembly tasks found that the VR group was able to complete the task almost three times faster than those traditionally trained (Oren et al., 2012). Furthermore, a manufacturing firm concluded that although VR training required more time, it led to an enhanced perception of one's ability to complete the task, resulting in a more confident worker who is less prone to mistakes (Smith & Salmon, 2017). However, VR can only be effective when the virtual environment closely matches the physical environment, necessitating good resolution and fidelity (Horejsi, 2015). In the military, the most prevalent application of VR technology can be seen in the Air Force Pilot Training Next Program (PTN), the predecessor to the Army's ATN program. Through low-cost VR simulators, the program's goal is to graduate pilots in only 14 weeks (Bolinger, 2019). However, despite the progress of the program, a lack of data collected on the program prevents conclusions from being drawn on its effectiveness.

Although VR has the potential to supplement pilot training, it's nature of training also offers disadvantages that can hurt pilot training if not handled properly. Some of these disadvantages include inducing adaptation and compensatory skills, poor motion cues, and lack of user motivation. The perception of danger and stress levels of pilot candidates may significantly differ between those using simulators and real aircraft. (Myers et al., 2018). Even though most of the simulators aim to perfectly recreate motion cues and avionics, they cannot encompass the real experience – stress and danger level – that real aircraft provides. However, a combination of real aircraft training and VR, as ATN is configured, will bridge the gap of training created

by the VR. Moreover, instructor pilot's knowledge of both VR and real aircrafts will enable to them to point out potential downfalls of VR to pilot candidates, hence improving the training transferring rate.

2.3 Presence Importance and Limitations

A VR system will always offer the same level of immersion, but the level of presence will vary because of a user's psychological response to the simulation (Bowman & McMahan, 2007). If immersion and presence are combined appropriately, the simulator will yield results that will produce acceptable results for similar tasks. To enhance VR experiences and results, mixed reality can also be incorporated which will allow for a user to learn faster and be immersed in a virtual and real-world experience for their task familiarization (Howard, 2019). VR interactions are dictated by many aspects which all contribute to a user being better prepared for real-life scenarios and tasks that VR systems try to model; therefore, these aspects should be incorporated into the ATN program for optimal results.

VR can have physical effects on the user. The University of Nottingham conducted a study that evaluated VR induced side effects. The study closely monitored the effects VR had on sickness, postural inability, psychomotor control, perceptual judgment, concentration, stress, and ergonomic effects (Cobb, 1999). Simulator sickness was among the highest of the physical effects that was felt following VR immersion. This was due to feedback from the visual system suggesting more movement than the brain interprets. A slower processor was correlated with high levels of disorientation in the users due to its greater lag of the update display (Cobb, p. 170). Additionally, VR has psychological effects on users. For example, VR based rehabilitation therapy results in more calm and relaxed patients. These patients, assessed in a hospital, felt decreased tension and increased calmness possibly because they didn't feel as though they were participating in rehabilitation therapy (Chen, 2009). The study concludes that the rehabilitation resulted in a benefited psychological state for patients due to calmness.

3. Data Analysis of Historical Data

To analyze the effectiveness of VR in the production of quality aviators, it is important to first establish a baseline standard for the qualities that define a good aviator. For the purpose of this study, nearly one year's worth of historical testing scores from graded assessments, both written and physical, were analyzed to determine average scores and performance. In coordination with decision making with USAACE senior leaders, the P1 and P2 checkrides emerged as the key determinants in the success of a flight school students within this effort's scope. Both checkrides are major milestones in a flight student's training, represent the final gate of each phase, and are highly standardized events that are easily compared across classes.

3.1 Methods of Analysis

The historical data used for this project's analysis contains five classes from April through September 2019. These classes were trained using the traditional program of instruction, known overall as Flight School XXI. The data consisted of flight school student scores for each graded event. There was a total of 232 data points across the entire dataset.

Beginning with analysis of the P1 checkride, the data was first used to determine whether there was a need for a control group in the ATN study. If the five traditional flight school classes demonstrated similar competency and performance, there would be no need for a control group to which to compare ATN scores. Instead, general historical data could be used as the control data for the study. A t-test was used to compare each class to each other with 95% confidence. The null hypothesis associated with this t-test was that there would be no statistical difference in the P1 and P2 checkride scores between each class. This would indicate that every class generally displays the same level of performance during the primary phase.

The next objective of the historical data analysis was the determination of general trends on the P1 and P2 checkrides. Specifically of interest were the average score, the standard deviation, the type and symmetry of the distribution, and the pass rate of the evaluations. By determining general metrics in these areas, a baseline to which to compare ATN data is established. If the ATN program can produce aviators of the same quality as traditional flight school means, the data from the ATN program can be expected to be very similar to the historical data.

3.2 Findings

After analyzing the P1 checkride scores across all five historical classes, there was found to be no significant differences in performance between classes. In the comparison of the classes, the t-tests failed to reject the null hypothesis and supported the idea that each class is similar regarding performance with each t-test producing a p value greater than 0.05.

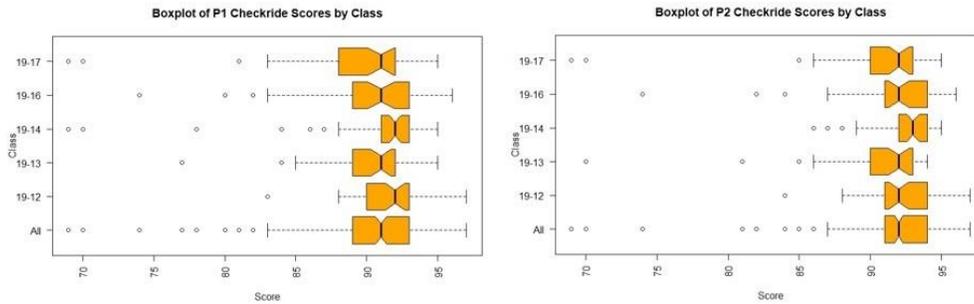


Figure 1. Boxplots of Historical P1 and P2 Checkride Scores

As further shown in Figure 1 with class numbers on the right and checkride score across the bottom, all five classes displayed similar score distributions and the average scores of all five classes were within one standard deviation of each of the other classes. All five classes showed a high success rate on the evaluation and each class had either zero or one failures (a score of 69 or lower). Therefore, based on the output of the t-test and the plots below, performance is independent of class and there is no need for a control group in this study. Instead, this historical data will be used as the baseline for what constitutes a quality aviator. Analysis of the historical P1 checkride score data indicated that students generally perform well on this event. The average score was approximately 90/100 with a standard deviation of about 4.5 points. There was also a large concentration of scores in the 90-93% range, further supporting the conclusion that this is an event in which students typically have done well. Lastly, across the five classes of date there were only three total failures.

The P2 checkride data illustrated similar conclusions with an average score of 91 and standard deviation of 3.9. Again, there were two failures and a high concentration of scores above the average. Like the P1 checkride, flight students generally do quite well on the P2 checkride. Data from the P1 and P2 checkrides in the ATN program should generally align with these conclusions and should be similarly distributed if the program is capable of producing aviators of similar quality. Indicators of ineffectiveness of the ATN program will be a rise in failures, a drop in the average score, or an increase in standard deviation.

4. ATN Program Data Analysis

4.1 Data Overview

Key to the evaluation of the success of the ATN program is the collection and analysis of relevant data that can indicate whether the program is capable of training capable aviators. The test scores were collected on every class that utilized the VR equipment and abided by the ATN curriculum beginning in September of 2019 through March of 2020. The data included all the scores that students in the ATN program earned on their written and practical evaluations. As before, the primary focus of the data analysis was on the P1 and P2 checkride, however, every test was analyzed for consistency with traditionally trained flight school students. Important to note, however, is that many of these classes were currently in session during the time of analysis and thus were incomplete as there were assessments not yet completed by multiple classes. This data was received as the classes progressed.

Another type of data used in the analysis of VR effectiveness was demographic and survey data provided by each student as part of a joint study with the US Army Aviation Aeromedical Research Lab. The demographic data was collected and analyzed with respect to their performance. Some of the demographic data used included but was not limited to gender, previous flight experience, video game experience, motion sickness history, and current stress level.

4.2 Methods of Analysis

Fort Rucker provided a comprehensive list of all students (identified by a survey identification number) that included their up-to-date test scores and extensive demographic data for each live and VR flight class. The datasets were filtered by class and included live aircraft flight scores and some demographic data that could possibly affect aircraft performance. Summary statistics for traditional and VR class data analyzed. The VR dataset contained 116 samples, whereas the traditional dataset had 145 samples. From the aggregate VR class, the P1 checkride and P2 checkride score averages and standard deviations were 89.8 and 3.7 and 91.7 and 3.9, respectively. Additionally, from the aggregate traditional class, their respective averages and standard deviation were 90.9 and 3.5 and 91.5 and 3.5. Figure 2 below shows two density plots overlaying histograms of the traditional and VR class flight evaluation scores.

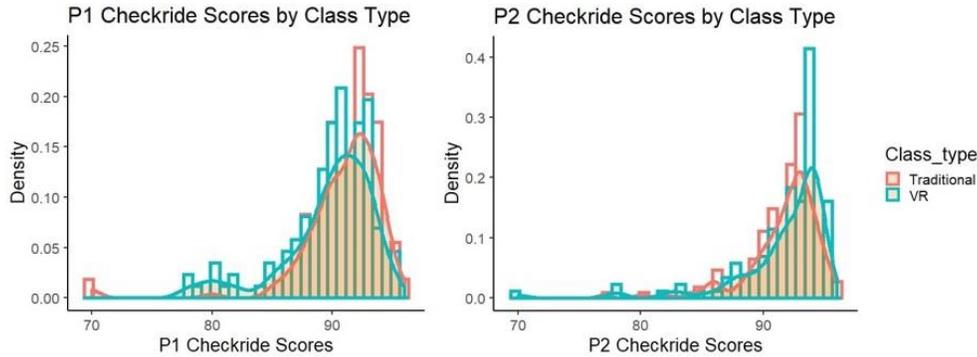


Figure 2. Density P1 and P2 Checkride Scores by Class Type

To analyze the possible demographic impact on flight evaluation, a linear model was developed using backwards regression. The P1 checkride score linear model yielded a 5.15% Adjusted R-Squared which supported that none of the demographic data related to determining aircraft performance. Similarly, the model utility hypothesis test supported none of these demographic variables determined aircraft performance. Since the P1 checkride score linear model yielded no statistically significant results, the other flight evaluation tests were not analyzed.

Since the VR program specifically tested the P1 checkride, a two-sample t-test was applied to examine the difference of the VR and traditional class mean. The aggregate traditional and VR class averages were statistically different with a p-value of 0.01221. Since the aggregate data yielded a statistical difference, the disaggregate data was also analyzed to give a more accurate representation of the true statistical difference between each class. The aggregate includes class differences that would skew data. For the 20-501/001, 20-504/004, and 20-505/005 classes there were no statistical differences in their averages with p-values of 0.7145, 0.6299, and 0.1668 respectively. Class 20-503/003 did have statistical difference in their averages with a p-value of 0.03977 which was likely the outlier class that skewed the aggregate data conclusion.

The ATN program only assesses skills and performance pertaining to the P1 checkride, but its results may extend to the P2 checkride. The same methodology for the P1 checkride analysis was done to analyze the difference of means between the classes. For the aggregate data, there was no statistical difference between the tradition and VR classes with a p-value of 0.6821. Additionally, the disaggregate data analysis supported the same statistical conclusion with classes 20-501/001, 20-503/003 20-504/004, and 20-505/005 yielding p-values of 0.01069, 0.5545, 0.6202, and 0.6601 respectively. This P2 analysis supports overall conclusion that the ATN program yields aviators of comparable skill proficiency as the traditional program.

4.3 Findings

The original summary statistics analysis supported the VR classes had a lesser standard deviation than the traditional classes. This was likely due to the VR classes being taught by Quality Assurance Instructor Pilots who would hold the students more closely to the standard due to additional training received to ensure more consistent subjective feedback. The linear model analysis supports that demographics play no role in determining an individual’s performance in an aircraft. This conclusion was likely skewed due to the lack of data across all classes and students. For example, it is likely that a pilot with over one hundred hours of fixed wing experience will do better on their flight evaluation scores, but since they would be an outlier in the dataset, a linear model would not reflect that fixed wing hours impact flight evaluation scores. Additionally, there was no statistical difference between the traditional and VR classes. Lastly, the average score for the P1 checkride in the ATN program is 89.8% and 91.7% for the P2 checkride, which almost exactly match the averages of the historical data. This analysis supports the conclusion that the ATN Program develops “just as good” aviators at Fort Rucker.

5. Conclusion and Future Work

5.1 Key Findings

Currently, it can be reasonably concluded from the data that the use of VR to supplement flight training does produce aviators of the same quality and competence as pilots trained in a live aircraft. As of April 2020, however, small sample sizes

mean that it is premature to draw definitive conclusions regarding ATN's effectiveness, despite early successes. Data collected through March of 2020 indicates that there is no significant drop in performance based on the P1 checkride and that there is potential for the program to consistently produce capable aviators of the same quality. Therefore, it is recommended that the program and study be continued to collect additional data in order to draw more significant conclusions and allow decision-makers more information when deciding whether to fully incorporate the ATN program into flight school.

5.2 Current Plan for VR at Fort Rucker

The current plan for the ATN program at Fort Rucker is to continue to implement it on a selective basis, as part of the ongoing design of experiments to determine the most effective combination of live to VR flight hours. Only a fraction of the overall students at Fort Rucker will be enrolled in ATN and these classes will be continued to be monitored and studied. This current plan facilitates the collection of more data to support further conclusions regarding its effectiveness. The plan is to keep manipulating the ratio of live and VR flight hours in the hope that the ideal ratio can be found and utilized. This study is ongoing, and the conclusions drawn in this paper are not final, rather working conclusions as the study progresses.

5.3 Future Work

It is necessary to continue to monitor and analyze data to create significant conclusions on the effectiveness of VR in flight school. As more classes progress through the ATN program, more data will become available to generate significant conclusions. There should be enough data by the Fall of 2020 to confidently conclude that VR whether is effective as more classes progress through the curriculum and the sample size grows. Furthermore, there will have been enough interactions to better determine the optimal ratio of live to VR flight hours in the curriculum. Further, another area of interest identified during this study was the ability to use VR systems as a means to assess a candidate's aptitude to be branched into the aviation community, with an objective of supplementing or replacing existing accessions tools that may not be as effective. Lastly, regardless of the success of the program, the lessons from this ATN program can be used and implemented into other Army training areas to supplement or enhance training.

An emerging area of exploration for this project involves cognitive training to enhance VR learning. The Cognitive Enhancement and Performance Program (CEPP) is an addition to the ATN, which aims to provide a holistic approach to optimize pilot's performance by fostering a strong learning environment, otherwise known as cognitive learning. CEPP targets the approach to instructions and learning relationships between instructors and students. Based on the analysis of effectiveness of CEPP in currently applied sections of Flight School, CEPP can extend its application of cognitive learning to other areas of Flight School. It is essential to evaluate the current effectiveness of CEPP through testing its measurements against students' performance to validate those measurements. Although not initially implemented synergistically with ATN, raw data from this initiative may allow for a subsequent redesign and improvements for both of these complementary programs.

6. References

- Bolinger, J. (2019, October 17). *Virtual Reality-Heavy Course Speeds Up Air Force Helicopter Pilot Training by Six Weeks*. Retrieved from Stars and Stripes.
- Bowman, D. A., & McMahan, R. P. (2007). Virtual Reality: How Much Immersion Is Enough? *IEEE*, 36-43.
- Chen, C. H., Jeng, M. C., Fung, C. P., Doong, J. L., & Chuang, T. Y. (2009). Psychological benefits of virtual reality for patients in rehabilitation therapy. *Journal of sport rehabilitation*, 18(2).
- Cobb, S. V., Nichols, S., Ramsey, A., & Wilson, J. R. (1999). Virtual reality-induced symptoms and effects (VRISE). *Presence: Teleoperators & Virtual Environments*, 8(2), 169-186.
- Committee on Developments in the Science of Learning. (2004). *How People Learn: Brain, Mind, Experience, and School*. Washington D.C.: National Academy Press.
- Horejsi, P. (2015). Augmented Reality System for Virtual Training of Parts Assembly. *Procedia Engineering*, 701-706.
- Howard, M. C. (2019). Virtual Reality Interventions for Personal Development: A Meta-Analysis of Hardware and Software. *Human-Computer Interaction*, 205-239.
- Myers, P. L., Starr, A. W., & Mullins, K. (2018). Flight Simulator Fidelity, Training Transfer, and the Role of Instructors in Optimizing Learning. *International Journal of Aviation, Aeronautics, and Aerospace*.
- Oren, M., Carlson, P. E., Gilbert, S. B., & Vance, J. M. (2012). Puzzle assembly training: Real world vs. virtual environment. *Mechanical Engineering Conference Presentations, Papers, and Proceedings*, 27-30.
- Smith, J. W., & Salmon, J. L. (2017). Development and Analysis of Virtual Reality Technician-Training Platform and Methods. *Interservice/Industry Training, Simulation, and Education Conference*, 1-12.