Pilot Training Transformation: Early Results and Lessons Learned

Samantha N. Emerson, Kent C. Halverson, Cait Rizzardo, Ramisha Knight, Julia Brown, Audrey Reinert

Aptima, Inc.

Woburn, MA

semerson@aptima.com, khalverson@aptima.com, crizzardo@aptima.com, rknight@aptima.com, jbrown@aptima.com, areinert@aptima.com

Mark G. Hoelscher, Tracy Schmidt, Lisa Tripp

United States Air Force, Air Education and Training Command
San Antonio, TX

mark.hoelscher.3@us.af.mil, tracy.schmidt@us.af.mil, lisa.tripp.1@us.af.mil

David Mills
The Perduco Group
San Antonio, TX
david.mills@linquest.com

ABSTRACT

The US Air Force (USAF) today faces a changing landscape in modern warfare, where pilots are expected to operate in highly technological environments requiring an increased amount of information synthesis and other skills commonly known as "airmanship." They will also be expected to adapt to fast-paced changes in technology and the accompanying strategy and tactical application. It was proposed that traditional training programs were not going to meet this need, and AETC set out to modernize its pilot training programs to better prepare the graduates by leveraging a variety of training simulation technologies such as Virtual Reality (VR) as well as modern learning methods to rebuild and modernize the Undergraduate Pilot Training (UPT) curricula. Although the performance standards for graduation had not changed, the training incorporated more dynamic training scenarios to further develop competency beyond basic proficiency. In particular, these curriculum changes were intended to focus on five objectives: (1) Enable seamless access to content, (2) transition the curricula to learner-centered training, (3) integrate immersive technology, (4) deliver quality instruction, and (5) optimize human performance. AETC senior leaders wanted additional, empirical evidence to determine if the new curricula were more effective than the legacy curricula as measured by performance of students from both programs. They chartered a study involving the development of controlled flight profiles and a battery of performance measures to capture a variety of performance subdimensions (e.g., mission planning, basic aircraft control, task management). The purpose of this study is to answer questions about the effectiveness of the revised curricula in the Pilot Training Transformation format and capture "lessons learned" that can inform a continuous improvement approach to UPT.

ABOUT THE AUTHORS

Dr. Samantha N. Emerson is a Scientist at Aptima, Inc. with over a decade of experience designing and executing rigorous research on human learning, thought, and language. She has a multidisciplinary background that combines the theories, methods, and analytics of psychology, cognition, neuroscience, and psycholinguistics. Prior to joining Aptima, Dr. Emerson served as a Postdoctoral Research Scientist at Boys Town National Research Hospital where she examined the learning and productive use of visual and auditory patterns in the environment. She holds a **PhD and MS in Cognitive Psychology** from Georgia State University and a BS in Psychology from Middle Tennessee State University.

Dr. Kent C. Halverson is a Principal Scientist and Senior Director of the Training, Learning, and Readiness Division at Aptima, Inc. He has research experience in performance measurement, organizational analysis, workflow modeling,

training design and evaluation, survey development, social network analysis, and quantitative data analysis. Dr. Halverson manages a portfolio of research projects for a variety of DoD research agencies such as the Air Force Research Lab, across a variety of operational domains (e.g., military pilots, infantry soldiers, intelligence analysts), levels of analysis (e.g., individuals, teams, organizations), and research perspectives (e.g., cognitive, behavioral, process, system). Dr. Halverson has held faculty positions at the Air Force Academy and the Air Force Institute of Technology as the Director of the Graduate Engineering Management Program. He holds a **PhD in Business Administration (Organizational Behavior)** from University of Florida, a MS in Civil/Structural Engineering from the University of Illinois, and a BS in Civil Engineering from the US Air Force Academy.

Dr. Cait Rizzardo is a Scientist at Aptima, Inc. with experience in performance assessment, evaluating simulation-based training programs, spatial navigation research, and in developing testing protocols for research done in virtual environments. Currently, she supports several projects researching training effectiveness across a variety of audiences including pilots, maintainers, and ground and logistics operators for both the Air Force Research Laboratory (AFRL) Warfighter Interactions & Readiness Division and Air Education and Training Command (AETC). Dr. Rizzardo has assisted with and led projects involving spatial knowledge acquisition, cross-cultural linguistic norms, display design and evaluation for after-action review and, prior to joining Aptima, Dr. Rizzardo conducted graduate research on spatial knowledge acquisition while using a redesigned GPS display, which resulted in an award-wining article in the *Journal of Experimental Psychology: Applied*. She also assisted in developing a spatial-acoustic navigation study with the AFRL Battlespace Acoustics Branch. Dr. Rizzardo holds a **PhD and MS in Human Factors/Industrial and Organization Psychology** from Wright State University and a BA in Psychology from Hollins University.

Dr. Ramisha Knight is a Scientist at Aptima, Inc. who specializes in Human Cognitive Neuroscience and Experimental Psychology. She has experience using a broad range of neurophysiological techniques to measure cognitive processes and the neural architecture associated with visual attention and perception. Her research includes statistical methods and machine learning-based approaches to predict behavior and performance. Prior to Aptima, Dr. Knight completed her postdoctoral work at the University of Illinois at Urbana-Champaign and the Beckman Institute for Advanced Science and Technology. She holds a **PhD in Psychological and Psychiatric Science** from Università degli studi di Verona (Italy), a **MS in Cognitive Neuroscience** from the University of Durham (England), and a BA in Psychology from Hawaii Pacific University.

Ms. Julia Brown is a Scientist in the Learning and Training Systems Division at Aptima, Inc. Ms. Brown has experience in selection, training, quantitative and qualitative analysis, and leadership development. She specializes in multi-rater, multi-method performance assessments. At Aptima, she currently supports a variety of projects to help military clients (a) develop requirements for new training environments, (b) evaluate new training curriculum, and (c) identify improvements to talent management systems. Prior to coming to Aptima, she worked as an external consultant contracting with Special Operations groups to measure performance in Assessment & Selection courses. Using research and knowledge elicitation methods, she analyzed job requirements in elite military organizations through focus groups, interviews, and on-site field observation. She has experience developing competency models, aggregating data, and designing reports to facilitate decision making. Her previous work also focused on the design and administration of peer assessments to provide feedback in military settings, as well as for D1 college athletes. Ms. Brown earned an **MS in Industrial/Organizational Psychology** from San Diego State University and a BS in Psychology from James Madison University.

Dr. Audrey Reinert is a Research Engineer in the Performance Augmentation Systems Division of Aptima. She holds a **PhD in Industrial Engineering** from Purdue University (2019), a **Masters in Human Computer Interaction** from the Georgia Institute of Technology (2015), and a Bachelors in Cognitive Neuropsychology from the University of California, San Diego (2012). Her research interests lie at the intersection of human centered computing, data visualization, and human machine interaction. Prior to joining Aptima, Dr. Reinert completed a postdoctoral work at the University of Oklahoma and Purdue University.

Dr. Mark G. Hoelscher is the technical advisor for Test and Training Resources in the US Air Force Air Education and Training Command, Studies and Analysis Squadron. Prior to joining government civil service, he was a Principal Research Scientist at the Georgia Tech Research Institute and the Faculty Research Leader for Live Virtual Constructive training. As a Senior Program Manager for SAIC, he led their involvement with the US Air Force's Pilot

Training Next – Advanced program. Dr. Hoelscher also served 26 years in the Air Force, retiring as a Colonel in 2017. During that time, he served as the Combined Test Director of the Joint Strike Fighter Operational Test Team, the Division Chief of the War Fighter Readiness Research Division in the Air Force Research Lab, and the Deputy Chief of the Electromagnetic Technology Division. He is a graduate of the US Air Force Test Pilot School with 2000+ flying hours in 36 aircraft types. Dr. Hoelscher holds a BS in Physics from the US Air Force Academy and a **PhD and MS in Applied Physics** from the Air Force Institute of Technology where he specialized in Optics and was awarded a US patent for remote sensing of hidden objects.

LtCol Tracy "Matrix" Schmidt is the Chief of Curriculum at 19th Air Force. She is an instructor and evaluator pilot in the USAF, with over 3,700 hours in the T-6A, T-38 and F-15E. In her current capacity, she oversees the Instructional Systems Design process for syllabus updates and new programs in development under the Pilot Training Transformation effort. She manages several contracts conducting ISD as well as developing a supporting Learning Management System. She also coordinates & integrates other specialties such as Human Performance Training under the Comprehensive Readiness for Aircrew Training (CRAFT) program, as well as data analysis efforts. Previously, as the Deputy Operations Group Commander at Vance AFB, she led the effort to incorporate PTN lessons learned into the UPT syllabus and conducted the first two test classes in 2019-2020, which became the basis for UPT 2.5. She has also deployed twice in the F-15E in support of OPERATIONS IRAQI & ENDURING FREEDOM.

Dr. Lisa Tripp is the Technical Director for the United States Air Force Air Education and Training Command, Studies and Analysis Squadron. The mission is to accelerate the transformation of Airmen development through relevant, responsive, and rigorous analytics. Prior to this position, Dr. Tripp worked for the Air Force Research Laboratory. Dr. Tripp has over 10 years of experience in the field of training research. Dr. Tripp holds a BS in Mathematics from Western Washington University, a MS in Applied Mathematics from Washington State University, and a PhD and MS in Experimental Psychology from Washington State University.

Dr. David Mills is lead support analyst for United States Air Force Air Education and Training Command, Studies and Analysis Squadron. He has held similar roles with other Air Force organizations and in non-military industry, both as an Active Duty US Air Force Officer and as a civilian. Dr. Mills retired from the US Air Force as a Major in 2014, and has over 25 years combined experience in cryptography, operational test and evaluation, mission effectiveness analysis, and academia. He holds a BS in Mathematics from Texas State University, a MS in Operations Research and a PhD in Applied Mathematics (Statistics), both from the Air Force Institute of Technology.

Pilot Training Transformation: Early Results and Lessons Learned

Samantha N. Emerson, Kent C. Halverson, Cait Rizzardo, Ramisha Knight, Julia Brown, Audrey Reinert
Aptima, Inc.

Woburn, MA
semerson@aptima.com, khalverson@aptima.com, crizzardo@aptima.com, rknight@aptima.com, jbrown@aptima.com, areinert@aptima.com

Mark G. Hoelscher, Tracy Schmidt, Lisa Tripp

United States Air Force, Air Education and Training Command
San Antonio, TX
mark.hoelscher.3@us.af.mil, tracy.schmidt@us.af.mil, lisa.tripp.1@us.af.mil

David Mills
The Perduco Group
San Antonio, TX
david.mills@linquest.com

INTRODUCTION

The US Air Force (USAF) today faces a changing landscape in modern warfare, where pilots are expected to operate in highly technological environments requiring an increased amount of information synthesis and other skills commonly known as "airmanship." They will also be expected to adapt to fast-paced changes in technology and the accompanying strategy and tactical application. It was proposed that traditional training programs were not going to meet this need, and AETC set out to modernize its pilot training programs to better prepare the graduates by leveraging a variety of training simulation technologies such as Virtual Reality (VR) as well as modern learning methods to rebuild and modernize the Undergraduate Pilot Training (UPT) curricula. Although the performance standards for graduation had not changed, the training incorporated more dynamic training scenarios to further develop competency beyond basic proficiency. In particular, these curriculum changes were intended to focus on five objectives: (1) Enable seamless access to content, (2) transition the curricula to learner-centered training, (3) integrate immersive technology, (4) deliver quality instruction, and (5) optimize human performance. AETC senior leaders wanted additional, empirical evidence that the new curricula were more effective than the legacy curricula as measured by performance of students from both programs. They chartered a study involving the development of controlled flight profiles and a battery of performance measures to capture a variety of performance subdimensions (e.g., mission planning, basic aircraft control, task management). The purpose of this study is to answer questions about the effectiveness of the Pilot Training Transformation (PTT) approaches and capture "lessons learned" that can inform a continuous improvement approach to UPT.

This study compared performance of recently graduated pilots who completed either the legacy curriculum (i.e., 2.0) or the revised curriculum (2.5). USAF UPT involves three phases, with the first focusing on academics and the second teaching students how to fly the T-6 Texan and allowing them to earn their pilot wings. For the third phase of training, students choose a track related to three primary classes of aircraft: fighters/bombers, airlift/tankers, and helicopters. Graduates selected to fly fighters/bombers next train on the T-38 Talon aircraft while graduates selected to fly airlift/tankers train on the T-1 Jayhawk aircraft. Curriculum version was always between both aircraft that the student was trained on (e.g., students who completed the 2.0 curriculum for the T-6 also completed the 2.0 curriculum for training in either the T-38 or T-1). There were comparatively few graduates selecting the helicopter track and therefore were not included in this evaluation. To gain a better understanding of the long-term effects of the new curriculum on student performance, we examined students after the second (T-6) and third phases (T-38 or T-1) of training. (See Meek, Hoelscher, Danley, & Brown, 2022 for a concurrent study being conducted on the T-38 and T-1 platforms as part of the PTT initiative.)

In contrast to previous examinations of pilot curriculum, which often focus on institutional assessment data collected while training in live aircraft, the present study uses data from scientifically developed flight profiles to be performed in a simulator. Profiles were designed by instructor pilots (IPs) to be sufficiently challenging and to include task saturated environments (e.g., emergency procedures, in-flight mission planning) that would have been impossible to script and standardize in live flight. Using flight simulators to evaluate student pilot performance is a valid approach that has been long supported (Bell & Waag, 1998; Hays, Jacobs, Prince, & Salas, 1992; Macchiarella, Arban, & Doherty, 2006). Student performance data were obtained using scientifically developed rating forms with easily observable behaviors. IPs were formally trained on use of the new rating forms that involved calibration sessions to improve inter rater reliability. This scientific approach (i.e., controlled flight profile, new rating form) was intended to generate sufficient variance in student performance data than exists in institutional training assessment data (i.e., gradebooks) which tend to be extremely negatively skewed with minimal variance.

METHODS

Participants

Participants included 150 Undergraduate Pilot Training (UPT) students stationed at either Vance Air Force Base (Enid, OK) or Randolph Air Force Base (San Antonio, TX) and enrolled in either the traditional UPT curriculum (2.0) or the revised curriculum (2.5). Table 1 shows participant demographic information grouped by training location, curriculum version, and aircraft. The evaluation described in this paper followed students through two of their training phases, thus, some participants are counted in more than one group. Less than 3 percent of participant data was incomplete due to external factors.

Table 1 Participant Group Enrollment

		Vance	Randolph	Overall
T-6				
2.0	Sample Size	51	0	51
	Age (yrs)	25.04		25.04
	Flight Experience (hrs)	145.5		145.5
2.5	Sample Size	72	28	100
	Age (yrs)	24.81	26	25.41
	Flight Experience (hrs)	199.8	136.28	336.08
T-1				
2.0	Sample Size	48	0	48
	Age (yrs)	25.9		25.9
	Flight Experience (hrs)	206.1		206.1
2.5	Sample Size	51	0	51
	Age (yrs)	26.19		26.19
	Flight Experience (hrs)	330		330
T-38				
2.0	Sample Size	21	0	21
	Age (yrs)	25.28		25.28
	Flight Experience (hrs)	178.6		178.6
2.5	Sample Size	30	0	30
	Age (yrs)	25.4		25.4
	Flight Experience (hrs)	258.9		258.9

Procedures

Upon graduating from their respective course, each student was scheduled in a simulator to fly the profile with an IP as evaluator. The student watched a video describing the purpose of the study. The IP provided a Non-Disclosure Document (NDD) that indicated students are prohibited from sharing any information about the profile to ensure all

students experienced the profile with no advanced knowledge. The student completed a survey instrument that collected information about demographics and prior flight experience.

Mission Planning and Brief

The *Mission Planning* assessment allowed the rating Instructor Pilot (IP) to record the cognitive frameworks used by students to plan their tasks and determine if mission planning requirements were met, which provided valuable context for the IP to assess piloting skills during the profile. Students were given 45 minutes to prepare for the T-6 simulator profile, 25 minutes to prepare for the T-1 simulator profile, and 30 minutes to prepare for the T-38 simulator profile. The allowable times were established by IPs based on how much time should be required to plan each respective profile. The student briefed their mission plan to the IP who assessed the student using the nine-item Mission Planning Assessment measure with a five-point Likert Scale (Strongly Disagree to Strongly Agree) and a N/A option. After the brief, the IP asked the student, "How confident do you feel that UPT prepared you for your next phase of training?" Response options were recorded on a five-point Likert Scale with anchors that included Not at all Confident, Slightly Confident, Moderately Confident, Confident, and Very Confident.

Simulated Flight Profiles

Participants completed a simulated flight profile. Each platform had a unique flight profile that was developed through knowledge elicitations with SMEs. The profiles were designed to be challenging for the students and included high task saturation during later segments on the flight that required students to demonstrate airmanship capabilities.

Each flight profile began with a Ground Ops phase when the student ran a modified preflight checklist. The profile was divided into "legs" related to waypoints along a route. Legs included three phases of flight, a) departure; b) en route, and c) approach. During the profile, students were expected to complete standard flying tasks (e.g., communicating with Air Traffic Control (ATC), leveling off at safe altitudes, managing radios, setting and maintaining course) in addition to maintaining basic aircraft control and prioritizing tasks appropriately. The profiles were designed to enable the assessment of eight airmanship capabilities: proactivity, adaptability, task management and prioritization, communication, risk management and decision making, situational awareness, general knowledge of complex systems, and information management.

Proficiency was observed and assessed by IPs during the profile. The instrument used a three-point rating scale including Ineffective, Moderate, and Effective, as well as an option for "Did not accomplish." This option was selected if the student did not attempt the task in question, either because the student should have performed task and did not or because completing the task would not have been appropriate considering their previous actions. For the former situation, IPs selected the "Did not accomplish" option and also rated the task as Ineffective. Proficiency assessments were completed for various tasks in each segment of the flight simulation. *Failure Analysis* was a count of the number of tasks IPs had rated as Ineffective.

Misprioritized Tasks were tasks executed earlier than what IPs consider an acceptable window of execution. This was a subjective assessment by the IP based heavily on the student's situation. IPs indicated many flight tasks can be accomplished in different orders, but there are some tasks that if not accomplished during limited windows could introduce unnecessary risk by displacing task(s) that could create future task saturation. For tasks considered to be misprioritized, the IP would record a check mark.

Late Tasks were tasks executed later than what IPs consider an acceptable window of execution. Similar to misprioritized tasks, late tasks were a subjective assessment by the IP based heavily on the student's situation. For tasks that fell into this category, the IP would record a check mark in the column labeled Too Late.

Task Management (TM) and *Basic Aircraft Control* (BA) were repeated measures that assessed these skills at multiple times during the three legs. As opposed to the specific task Proficiency items, these repeated measures were intended to capture more generally how a student was performing since the last time they were rated on that item. Task Management refers to how well they were managing the numerous tasks they needed to complete. Task Management was not assessed separately in the T-38 platform. Instead, IPs were instructed to incorporate "task management" as a component in the ratings for all items. Basic Aircraft refers to their flying skills (e.g., maintaining a glide path). Scores obtained from these repeated measures were aggregated within each leg and across all three legs.

In the earlier phase of training on the T-6 platform, difficulty was manipulated across the different legs of the flight. In order to capture performance as a function of difficulty, periods with high task loads were identified. These *Task Saturation Events* were assessed for each leg by aggregating the Proficiency value for each of these high load. During the later phase of training (i.e., T-1, T-38) after students' abilities were more advanced, profiles were designed to involve a consistently high level of task saturation for the duration of the profile. As such, Task Saturation Events were not assessed in the T-1 or T-38 platforms.

Crew Management was assessed in the T-1 platform. The purpose of this repeated item was to assess how the student communicated with and utilized crew resources, including the copilot. This measure was not relevant to the T-6 or T-1 platforms, which are single seat aircraft.

Post-Mission Assessments

After the student completed the simulator profile and presented a debrief to the IP, the IP completed a *Post Mission Assessment* where they reflected on the student performance across the entire profile to make cumulative assessments. An example of an item is, "Did they send communications effectively?" Response options varied for each item but followed a general format: A. <50% of the time, B. 50-80% of the time, C. 80+% of the time.

Analyses

Statistical analyses were performed using SciPy in Python 3.7. Because observations were not normally distributed, parametric statistical analyses were not appropriate for this dataset. Instead, student performance comparing the legacy 2.0 versus the revised 2.5 curricula were assessed using a between-subjects Wilcoxon Mann-Whitney U test. Comparisons were considered statistically significant when their p-value was less than 0.05.

RESULTS

The following section contains three tables showing a comparison of the performance of pilots who received the legacy 2.0 syllabus with those who received the 2.5 syllabus.

T-6

The primary objective of the study was to determine if the T-6 2.5 curriculum resulted in student pilot performance that was equal to, or above, pilot proficiency acquired using the 2.0 curriculum. Table 1 and Figure 1 provide summary statistics for student pilot performance across nine aggregate performance metrics. Two-tailed Mann-Whitney tests indicated students who received the 2.5 curriculum had better ratings than the 2.0 students across all metrics, with statistically significant differences on mission planning ($U = 1960.5$, $p = 0.031$), basic aircraft control ($U = 1640.5$, $p < 0.001$), task management ($U = 1370$, $p < 0.001$), proficiency ($U = 1430.5$, $p < 0.001$), post mission assessment ($U = 1434$, $p < 0.001$), and task saturation ($U = 1164$, $p < 0.001$). Similarly, 2.5 students executed fewer tasks ineffectively than 2.0 students ($U = 3451.5$, $p < 0.001$) and made fewer timing errors ($U = 3585$, $p < 0.001$). The 2.5 curriculum students also made fewer mis-prioritized errors than 2.0 students, however, this result was not statistically different ($U = 2825$, $p = 0.255$).

Table 1 Comparison of students in the 2.0 versus 2.5 curriculum for the T-6.

Performance Metric	2.0 (mean)	2.5 (mean)	2.5 Change	*p*
Mission Planning (6pt Likert scale)	3.06	3.45	0.39	0.031
Basic Aircraft Control (3pt effectiveness scale)	2.45	2.63	0.18	<0.001
Task Management (3pt effectiveness scale)	2.37	2.63	0.26	<0.001
Misprioritized Tasks (count of misprioritized tasks*)	3.53	2.01	-1.52	0.255
Late Tasks (count of late tasks*)	8.51	3.85	-4.66	<0.001
Proficiency (3pt effectiveness scale)	2.71	2.81	0.10	<0.001
Post Mission Assessment (3pt frequency scale)	2.13	2.47	0.34	<0.001
Task Saturation Events (3pt effectiveness scale)	2.50	2.70	0.20	<0.001
Failure Analysis (count of ineffective tasks*)	14.00	8.52	-5.48	<0.001

* Lower scores = better performance; dark green = 2.5 significantly better than 2.0; light green = 2.5 nominally better than 2.0; dark red = 2.5 significantly worse than 2.0; light red = 2.5 nominally worse than 2.0

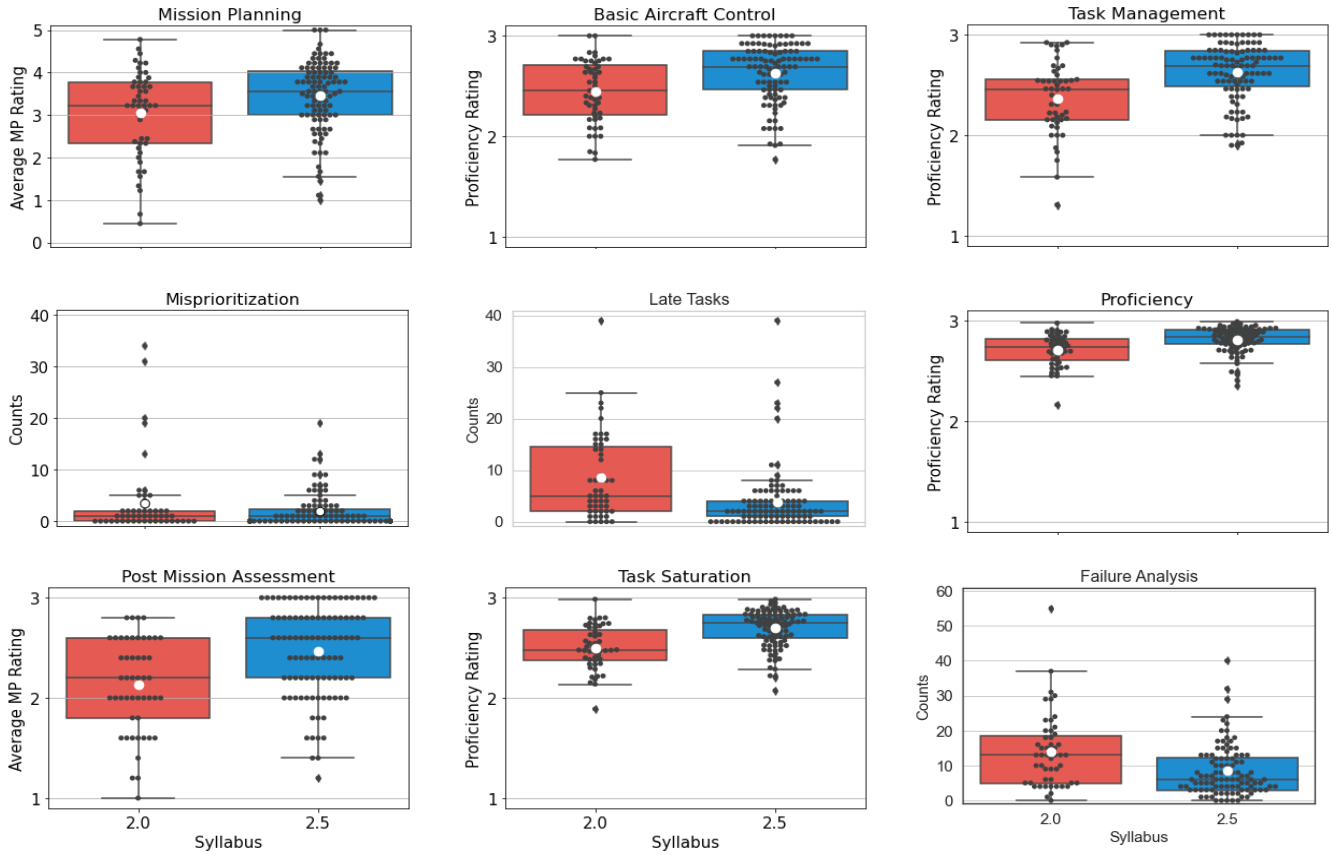

Figure 1. Box and whisker plots of T-6 legacy 2.0 (red) and revised 2.5 (blue) curricula.

T-1

Table 2 and Figure 2 compare of students who of the legacy 2.0 syllabus during T-1 training and those of the Air Mobility Fundamentals-Flying (AMF-F) version of the 2.5 syllabus. A Mann-Whitney test indicated students who received the 2.5 curriculum had significantly fewer late tasks ($U = 1576$, $p = 0.011$) than those who received the 2.0 curriculum. Further, the students of the 2.5 curriculum misprioritized fewer tasks ($U = 1621$, $p < 0.001$) than those of the 2.0 curriculum. There were no significant differences between the two groups on the remaining metrics.

Table 2 Comparison of students in the 2.0 versus 2.5 curriculum for the T-1 AMF-F.

Performance Metric	2.0 (mean)	AMF-F (mean)	AMF-F Change	*p*
Mission Planning (6pt Likert scale)	3.71	3.71	0.00	0.939
Basic Aircraft Control (3pt effectiveness scale)	2.58	2.57	-0.01	0.693
Task Management (3pt effectiveness scale)	2.61	2.54	-0.07	0.214
Crew Management (3pt effectiveness scale)	2.68	2.69	0.01	0.920
Misprioritized Tasks (count of misprioritized tasks*)	1.00	0.25	-0.75	<0.001
Late Tasks (count of late tasks*)	2.50	1.37	-1.13	0.011
Proficiency (3pt effectiveness scale)	2.60	2.59	-0.01	0.575
Post Mission Assessment (3pt frequency scale)	2.47	2.42	-0.05	0.831
Failure Analysis (count of ineffective tasks*)	13.77	15.02	1.25	0.402

* Lower scores = better performance; dark green = 2.5 significantly better than 2.0; light green = 2.5 nominally better than 2.0; dark red = 2.5 significantly worse than 2.0; light red = 2.5 nominally worse than 2.0

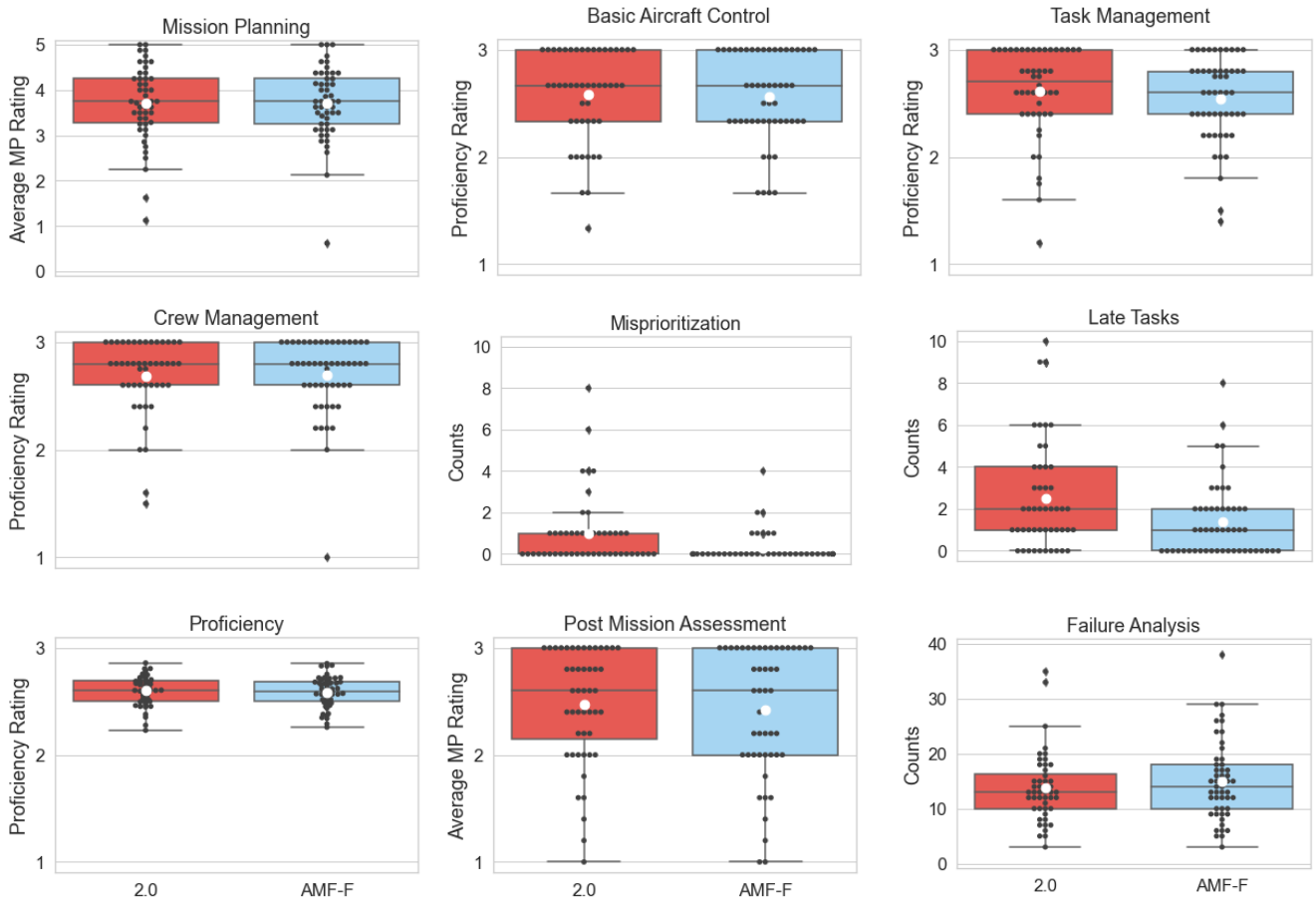

Figure 2. Box and whisker plots of T-1 legacy 2.0 (red) and revised AMF-F (light blue) curricula.

T-38

Table 3 and Figure 3 compare students who received the legacy 2.0 syllabus during T-38 training and those who received the 2.5 syllabus. Mann-Whitney tests indicated there were no significant differences between the two groups on any performance metrics.

Table 3 Comparison of students in the 2.0 versus 2.5 curriculum for the T-38.

Performance Metric	2.0 (mean)	2.5 (mean)	2.5 Change	*p*
Mission Planning (6pt Likert scale)	3.95	4.18	0.23	0.313
Basic Aircraft Control (3pt effectiveness scale)	2.58	2.54	-0.04	0.723
Misprioritized Tasks (count of misprioritized tasks*)	4.19	3.77	-0.42	0.496
Late Tasks (count of late tasks*)	5.95	6.87	0.92	0.463
Proficiency (3pt effectiveness scale)	2.83	2.82	-0.01	0.947
Post Mission Assessment (3pt frequency scale)	2.40	2.61	0.21	0.085
Failure Analysis (count of ineffective tasks*)	6.71	7.57	0.86	0.908

* Lower scores = better performance; dark green = 2.5 significantly better than 2.0; light green = 2.5 nominally better than 2.0; dark red = 2.5 significantly worse than 2.0; light red = 2.5 nominally worse than 2.0

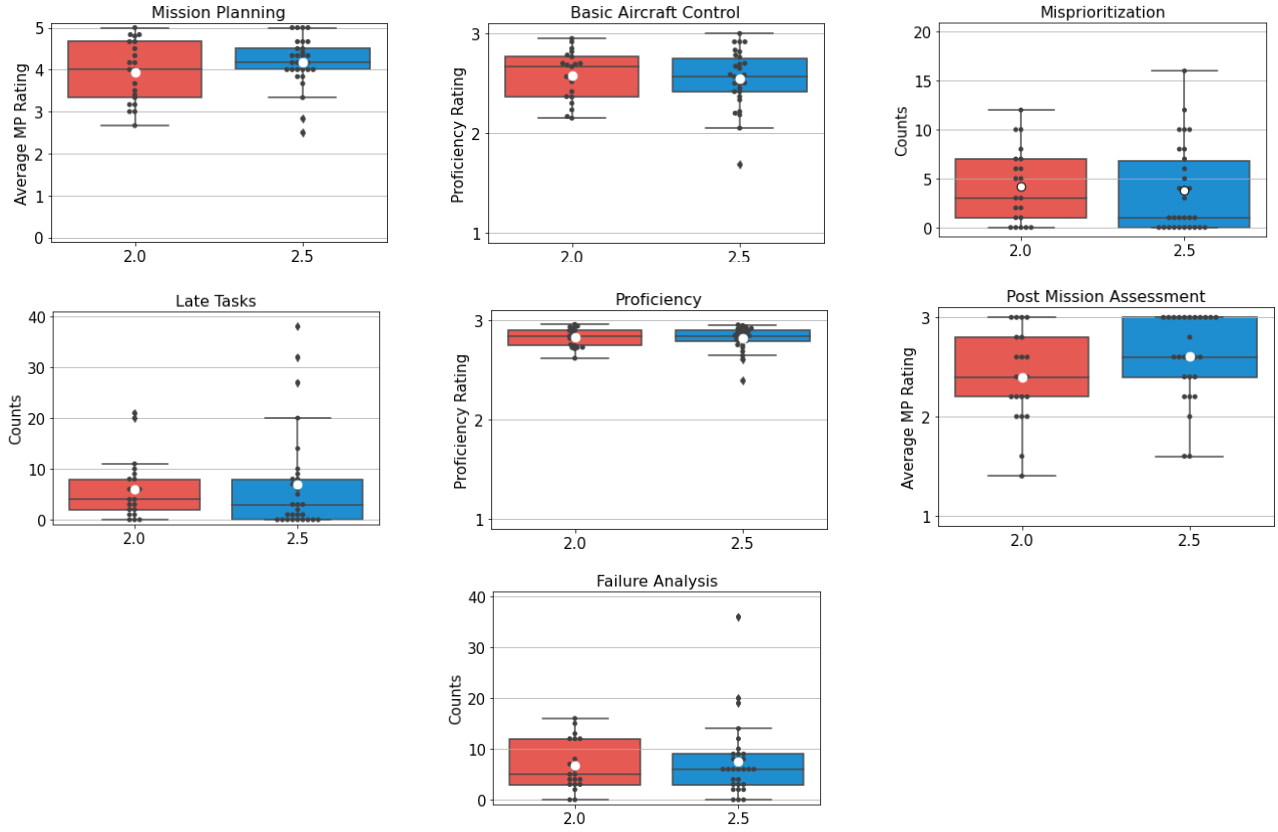

Figure 3. Box and whisker plots of T-38 legacy 2.0 (red) and revised 2.5 (blue) curricula.

DISCUSSION

Overall, the results of the study provide evidence the revised 2.5 curricula, relative to the legacy 2.0 curricula, produced pilots who performed at an equal or higher ability than their peers.

The most notable differences between the graduates of the two curricula were seen in the T-6 platform—the earlier phase of Undergraduate Pilot Training (UPT). Students of the 2.5 curriculum outperformed students of the 2.0 curriculum on all but one outcome metric, misprioritization. Interestingly, while performance on misprioritization did not differ between students of the two curriculums, graduates of the 2.5 curriculum did perform nominally (but not

significantly) better than graduates of 2.0 curriculum. These results could suggest the 2.5 curriculum was effective in improving the quality of UPT graduates on the T-6 platform.

Benefits of the new curriculum were less pronounced in the later phase of training (e.g., as students began to specialize in either the T-1 or T-38 platform). For the T-1 students, 2.5 graduates misprioritized significantly fewer tasks and were late on fewer tasks than the 2.0 graduates. For most other tasks, student performance was nearly identical. The one area where 2.5 graduates performed nominally (but not significantly) worse than 2.0 graduates was on the number of ineffective tasks; however, the difference amounted to an average of only 1.25 more items as being rated as ineffective compared to graduates of the 2.0 curriculum. It is worth noting that the flight hours in 2.5 curriculum for the T-1 was half of that of the 2.0 curriculum, having been re-focused on the most important unique aspects of T-1 flight (e.g., crew coordination, decision making). As such, the results of the analyses suggest that for the T-1, 2.5 graduates were trained up to an equal performance level as the 2.0 graduates, but in less time.

For the T-38 platform, 2.5 graduates performed at an equally high level as 2.0 graduates on all outcome metrics. Similar to results from the T-1 students, the T-38 2.5 graduates performed nominally (but not significantly) worse, having a greater number of ineffective tasks relative to 2.0 graduates. 2.5 graduates also produced a nominally greater (but not significantly) number of late tasks than their 2.0 graduates. Again, in both cases, the differences were minimal with less than a single item difference between the two groups. Thus, the results for the revised curricula for both the T-1 and T-38 produced graduates at similarly proficiency levels as the 2.0 graduates.

Recommendations for Further Curriculum Comparisons

Overall, results of the study suggest the 2.5 curricula were highly effective in producing graduates of an equal or higher caliber as the 2.0 curricula. In the later phases of training, three areas were noted where performance was nominally worse for graduates of the 2.5 curricula compared to the 2.0 curricula (i.e., T-1 Failure Analysis, T-38 Late Tasks, T-38 Failure Analysis). Future examinations of the revised curricula could probe these analyses for potential areas of improvement. However, it is important to note that because neither the overall analysis nor any of the analyses for the individual items were significant, any group differences may reflect statistical noise rather than actual effects. As such, we refrain from making suggestions for any substantive changes to the 2.5 curricula.

With respect to future studies, the data collection required a high level of effort from Instructor Pilots, who performed rater duties in conjunction with completing their normal daily roles and responsibilities. Instructor expertise is vital when developing an evaluative scenario containing events that cue key pilot behaviors. These behaviors demonstrate student skills and knowledge absorbed during their training, and those outcomes support determinations of training effectiveness. Rating these behaviors while students are in the simulator, often across multiple dimensions and for the full duration of a simulated scenario, is a mentally strenuous and fatiguing task. In future curriculum evaluations, the inclusion of supplemental system-based measures (e.g., Air Force Research Lab's Performance Evaluation and Tracking System; PETS) of student behavior, built into training devices and developed in accordance with defined training objectives, could provide reliable and meaningful objective performance data while reducing the demand on IP resources.

Conclusions

In this study, we examined student flight performance between graduates of the traditional legacy UPT curricula (2.0) compared to a revised curricula (2.5) which included a number of changes unique to each phase of training. Results revealed the revised curriculum had dramatic effect on student performance during the earlier phase of training (i.e., on the T-6 platform), where graduates of the 2.5 curriculum outperformed graduates of the 2.0 curriculum on nearly every outcome measure. In the later phase of training on the T-1 or T-38 platform, students in the 2.5 curricula were shown to have an equally high quality of performance as 2.0 graduates across most metrics. In no case did the 2.5 graduates perform significantly worse than the 2.0 graduates. As such, results provide strong evidence the revised curricula produce graduates of a high caliber.

REFERENCES

Bell, H.H. & Waag, W. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology*, *8*(3), 223-242.

Hays, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military psychology*, *4*(2), 63-74.

Macchiarella, N. D., Arban, P. K., & Doherty, S. M. (2006). Transfer of training from flight training devices to flight for ab-initio pilots. *International Journal of Applied Aviation Studies*, *6*(2), 299.

Meek, C., Hoelscher, M. G., Danley, T., & Brown, Q. (2022). *Comparing immersive versus legacy aviation training: A method for assessing student pilot proficiency*. Manuscript in preparation.