January 2013

Effects Of Instructor Continuity On A Large-Scale Pilot Training Program

Erik Michael Goff

EFFECTS OF INSTRUCTOR CONTINUITY ON A LARGE-SCALE PILOT TRAINING PROGRAM

by

Erik Michael Goff

Bachelor of Science, United States Air Force Academy, 2005

A thesis

Submitted to the Graduate Faculty

of the

University of North Dakota

In partial fulfillment of the requirements

for the degree of

Master of Science in Applied Economics

Grand Forks, North Dakota

May

2013

This thesis, submitted by Erik Goff in partial fulfillment of the requirements for the Degree of Master of Applied Economics from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done, and is hereby approved.

Dr. Cullen Goenner

Dr. Daniel Biederman

Dr. David Flynn

This thesis is being submitted by the appointed advisory committee as having met all of the requirements of the Graduate School at the University of North Dakota and is hereby approved.

Wayne Swisher
Dean of the Graduate School

19 April 2013

Title	Effects of Instructor Continuity on a Large-Scale Pilot Training Program
Department	Department of Economics
Degree	Masters of Science in Applied Economics

In presenting this thesis in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my thesis work or, in his absence, by the Chairperson of the department or the dean of the Graduate School. It is understood that any copying or publication or other use of this thesis or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my thesis.

Erik Goff
27 March 2013

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to my wife Audrey for all the support she gave me throughout this degree, and my Mother and Father, for showing me the value of education. I would also like to thank the members of my advisory committee for their guidance and support during my time in the master's program at the University of North Dakota.

ABSTRACT

In a large-scale pilot training program, like those run by the United States Air Force, Navy or many civilian colleges, quality of instruction is a very important question. These programs will have large numbers of instructors and students, but the effect of flying a student with a large number of different instructors is unknown. This paper will use continuous performance and failure data to see if there is a relationship between flying with too many instructors and failing or scoring poorly on a checkride. Multiple regression, difference in differencing, and probit analysis will be used to explore this question.

CHAPTER I

INTRODUCTION

Flight training has over 100 years of history, but when compared with other types of education it is still quite new. Many different types of training have been attempted, but there is still much more to learn about how to train pilots. Innovations have come through trial and error, technology, and best practices. One popular idea is "instructor continuity," or limiting the number of instructors for each student. Everyone has heard that "too many cooks will spoil the broth," but we do not know how applicable this axiom is to flight training. Most other forms of education adhere to this rule, but this is likely because it costs less to have just one teacher. A collection of case studies by Hall (1971) studied how young students are affected by having a multi-year teacher in Montessori, Kibbutzim or Orthogenic schools. The program discussed here is very different from those schools, but the inherent ideas are the same; the study talks about how a longer term student-teacher relationship builds mutual respect and increases the expectations on both parties.

This continuity axiom has been applied to United States Air Force flight training for many years. Only more recently have the flight training syllabi allowed for students to fly with larger numbers of instructors. The axiom may not apply to certain forms of flight training for a few different reasons. The skills associated with basic flight training are often not difficult to explain and demonstrate by a seasoned instructor, but they do require lots of practice by the student. Research on the topic divides the skills into two main categories: tracking and procedural. A tracking task is continuous, for example maintaining a specific altitude and airspeed, and a procedural task is discrete, like running a checklist or an emergency procedure (Gardlin, 1972). So unlike a liberal art, early flight training does not require a whole lot of creativity; there are extensive rule books and even technique books, and each USAF flight training unit has a rigorous standardization program so that the lessons should be fairly similar even with different instructors. For these reasons it may be okay to have more cooks in this particular kitchen. The purpose of this paper will be to evaluate the impact that the number of instructors has on student performance.

The Air Force, Navy and Army have several large scale pilot training operations that could certainly benefit from this research, but some college aviation programs such as North Dakota and Embry-Riddle are also large enough to benefit from greater scheduling flexibility. Data from a 2011 survey by Christiansen indicate that there are at least 16 collegiate flight training programs with over 10 aircraft, and 21 programs which have between 50 and 149 students at any one time. With these numbers of students and aircraft it is likely that increased scheduling flexibility could help streamline these programs.

The reason that pilot quality is especially important to the Armed Forces of the United States is that the student pilots could be flying in combat within 4-6 months of graduating from pilot training. Two of the largest end-users of T-1A students are the C-17A and KC-135R communities, whose training programs last three and six months respectively. After that training most new pilots will accomplish a month-long local qualification at their new unit, and the next stop is often a combat deployment. The student will be paired with an experienced aircraft commander, but the expectation is that the individual could also command the aircraft in the event of an emergency.

The cost of each military aircraft also reinforces the importance of training the people operating them. A report from the Government Accountability Office's 1997 Defense Aircraft Investments Report states "DOD's current aircraft investment strategy involves the purchase or significant modification of at least 8,499 aircraft in 17 aircraft programs at a total procurement cost of $334.8 billion." Further estimates from this report put the unit cost of a C-17A at $220 million (GAO, 1997, 2). Figures of this magnitude demand competent pilots at the controls, whether they are right out of training or a seasoned veteran.

With this greater understanding of the reasons why pilot training is important, it is easy to see that any improvement that could reduce costs or increase quality would be worth researching. This research will study scheduling flexibility; we will ask if we are producing better or worse pilots by allowing students to fly with more instructors in a particular phase of USAF Pilot Training. In flight training we talk about different types of continuity, for example instructor continuity refers to how often a student flies with the same instructor. Flying continuity will refer to how smooth the flight schedule was for any particular student or class of students. For example, most students would like to fly two to three times per week, this way they have time to plan before each flight, but they don't get "cold hands" during their time off. This concept of "cold hands" is confirmed in a study by Wilson (1973), which showed the effects of too much time off, and how low-skilled pilots are especially susceptible. The other form of continuity is instruction continuity which refers to how standardized the actual instruction and techniques were from one instructor to the next.

Other fields such as psychology have written extensively about learning and how to optimize learning a specific task. Pilot Training involves many different tasks; so many different types of research could apply. Memorization, synthesizing information, hand-eye coordination, critical thinking, and prioritization are a few of the skill types required of a pilot. It is difficult to study so many types of skills at once, so many of the smaller tasks have been studied separately. Savion-Lemieux and Penhume (1994) studied at how delays before recall affected the accuracy of recall in a timed motor sequence task (TMST). Their research showed that as time passed, the accuracy with which a subject performed a simple recall task dropped. They also showed that breaks between practice sessions improved task recall. The skills they tested were very basic, consisting of sequenced tapping, but nevertheless show that a good flow in a training program is exceedingly important.

One researcher, who studied a more complicated task, studied recall on the task of landing a jet aircraft on an aircraft carrier. Wilson (1973) took pilots with varying total flight hours and varying time out of the cockpit and analyzed their recall for one of the most difficult tasks in the aviation world. He found that total experience in piloting tasks was more significant than time out of the cockpit. This has interesting ramifications for more experienced pilots, but for inexperienced pilots it basically just indicates that more practice is better.

The importance of a landing task is obvious; all pilots main goal is to make sure their total takeoffs equal their total landings. Pilot training grades a large number of tasks, and the totality of these tasks is what will be studied in this paper. The goal will be to define how students are affected by larger numbers of instructors, by showing whether or not the two statistics are significantly related, and to see the magnitude of these effects. This will allow high and low level decision makers something to base their decisions upon.

Literature Review

The topic of pilot training has been written on extensively, the Congressional Budget Office sums up the issues in its 1994 report on pilot retention saying:

"The skills of military pilots are obviously essential to any mission employing combat air forces. Moreover, the role of pilots may grow in importance as the United States increases its use of air power in global peacekeeping missions. However, military pilot training is expensive." (Jehn, 1999, 1)

These three sentences succinctly state the forces at play, and emphasize the importance of pilot training as we increase our use of technology in the military. The previous research on pilot training will illustrate the importance of this topic, and validate the research methodology.

One of the best indications on the importance of this topic is the amount of research already completed on pilot training. Papers written on pilot training tend to follow a few common threads. Many of these papers concentrate on how to select candidates for pilot training and the effectiveness of simulator training. A meta-analysis written in 1992 surveyed over 40 different papers written on flight training, but those papers were mainly on different aspects of pilot-candidate selection. Factors like prior education, scores on aptitude tests and personality were used to predict performance. Success in pilot training is an important factor because according to this meta-analysis by Lynch each failure in pilot training cost the government between $65,000 and $80,000 (1992). Another cost analysis paper by D.J. Blower (1997) puts the costs of a training failure in naval pilot training as between $18,000 in the initial academic phase to as much as $500,000 in the advanced strike program. These papers do not offer methods for studying this particular research question, but they certainly show the importance of this topic due to the breadth of research and monetary involvement in the flight training industry.

One of the research papers that highlight the importance of studying pilot training is the research paper by Frizinger (1980) on economies of scale at work in USAF pilot training. Over the period of this study the number of USAF pilot training bases varied between 5 and 10, and total students graduated varied between 1000 and 4000. This paper is similar to research herein in that it seeks to study waste and efficiency in the pilot training system. Their question was about the ideal size of a pilot training base, while the question here is about how specific portions of the pilot training syllabus are executed. That particular study found that there were economies of scale present, so it would be useful to consolidate pilot training into fewer larger bases. Whether this research affected events or not, there are even less pilot training bases now some 30 years later.

A more recent paper from the University of North Dakota's aviation program (Schumacher, 2007) studied the difference in presenting pilot skills in a maneuvers-based format versus a scenario-based format. The study used a control (maneuver-based) group and a group using scenario-based training, where events are practiced as they would occur in a normal flight. This study then checked the student's maneuver scores and their scores on their stage checks (Schumacher, 2007) or what we would call checkrides in the USAF. The methodology of comparing the maneuver and check scores will be the same way that we gauge student performance in this paper.

The UND study talked about a teaching method used in training pilots, while a similar paper by Rokicki (1981) studied a different internal factor of USAF pilot training. This paper studied how the time that students showed up affected their perceived alertness. In this paper students and instructors filled out short survey cards telling when they went to sleep and woke up and if they felt well rested. This paper was one of the few papers that looked at a specific aspect of pilot training and studied whether a change could improve the value of training.

The research by Rokicki was different in that it was subjective, and not based on actual performance data from the student's grades. This research will use actual performance data to study an internal factor of pilot training in an effort to provide concrete data on how students are affected by instructor continuity.

Program Review

To fully understand some of the terms in this paper, and to better illustrate why this is an important research issue will require a small amount of background on the United States Air Force's Specialized Undergraduate Pilot Training (SUPT). SUPT is approximately one year long, and is one of the more intensive courses of pilot training available. It is divided into three phases: Phase I is academics, Phase II is primary T-6 Training and Phase III is either T-1 (tanker/transport) or T-38 (fighter/bomber) training. We will concentrate on the T-1 portion of Phase III of SUPT, which teaches the aerial refueling and airlift pilots.

The T-1 program lasts around 6 months. The first three weeks consist of academics and simulator training. After the initial portion, academics continue at a slower pace and students begin the flying portion of the program. Each student accomplishes 44 or more flying events (AETC Syllabus P-V4A-G, 2011), logging around 85 primary hours. The program is divided into 3 blocks of training: Transition, Navigation and Introduction to Mobility Fundamentals (IMF). Transition concentrates on basic aircraft control and the traffic pattern. Navigation focuses on visual, high and low level navigation, with emphasis on instrument approaches and the Instrument Flight Rules (IFR) system. The IMF phase introduces the students to heavy formation, air drop and simulated aerial refueling.

The goal of this program is to produce pilots (AETC Syllabus P-V4A-G, 2011), and results can be difficult to measure. Success at follow-on training is important, but also difficult to measure, because it happens at different bases all over the country. Performance on checkrides is a good measure because when a student fails a checkride they must fly another flight called a “progress check,” and then if they fail the progress check, they fly another flight called an “elimination check.” If the student fails an elimination check he or she goes under review to be removed from the program (AETC Syllabus P-V4A-G, 2011). This is important because the two extra flights would last around four flight hours, which would cost approximately $4200 in commercial aviation fuel (based on typical T-1 burn rate of 180 gallons per hour, times 4 hours, at $6 per gallon from Airnav, 2012) . Better checkride scores should decrease the number of extra flights that would be necessary, cut costs, and produce better pilots.

In this paper we will specifically study the 48th Flight Training Squadron (FTS) at Columbus Air Force Base (AFB), Mississippi. Columbus AFB is located about 10 miles north of Columbus, Mississippi, a town of about 24,000. In the early 1940's the leading citizens of Columbus banded together to lobby for an air base, offering to lease the land to the government for $1 per year. Based on the good location, cheap real estate, and the wide open airspace, Columbus AFB has thrived over its 70 year history. During the peak years of World War II, Columbus graduated around 195 pilots per month (Columbus History, 2011), which is close to the amount of pilots who now graduate per year. Columbus is currently one of three pilot training bases employing the T-1 for tanker/transport training, so the results of any research could affect around 300 instructors and around 1,500 students per year (Columbus, 2013 and Vance, 2011).

The 48 FTS has within its Operating Instructions (OIs) rules about how and when to limit students to a fewer number of instructors. This is likely hold-over from past syllabi where there was a requirement to limit how many instructors with which a student flew. Mr. Steve Babcock, the Point of Contact for the last three syllabi stated in an email on the 19th of October 2012 that continuity has not been specified in the syllabus since at least 1997, but that when he was an instructor pilot, continuity was specified in the pilot training regulations. Currently, the syllabus only states that flight commanders will ensure continuity, with no specifics on how this is accomplished

(AETC Syllabus P-V4A-G, 2011). Struggling students can also be limited to experienced instructors, but the squadron has many experienced instructors, so this is not difficult. These rules provide a framework, but the schedulers and flight commanders enforce these rules, and changes to squadron scheduling during the winter of 2012 made these rules more difficult to implement and increased the variability in this dataset.

CHAPTER II.

THE DATA

Before we delve into the reason that this is worthwhile research, an explanation of the variables is in order. The two main variable groups of interest are: student performance indicators and instructors per student or “counters.” Student performance data is collected from the Training Integration Management System (TIMS), a powerful computer system developed by Northrop-Grumman which collects and records training data for various military flight training programs. This dataset was taken from 144 students over 10 classes, all scheduled to graduate in Fiscal Year 2012. Data from all classes in this period was not collected in order to keep the data anonymous. Names and class numbers have also been replaced for the sake of privacy. An observation number and class identification number replace this data. Privacy is required because some of this information is never released, but the data is collected in a format where a student armed with his personal scores, could not identify him or herself in the dataset.

Performance was recorded in a few different ways. First were checkrides, which are final “test” flights following all the instructional flights of each phase. If we compare to a traditional school, checkrides would correspond with midterm and final exams. Other elements such as weekly quizzes, daily flight scores, or officership would fall in the daily maneuver grades or flight commander ranking categories. A dummy variable representing checkride failures was recorded along with raw checkride scores, which were recorded as a continuous variable representing the ratio of points scored over points attempted. Each maneuver that was attempted is worth a set amount of points (normally 5). The values correspond to the following grades: No Grade (if the maneuver was demonstrated), Unsatisfactory, Fair, Good and Excellent. These grades are clearly defined in the course syllabus. For example the grade Fair is defined as "The student performs the operation, maneuver, or task safely but has limited proficiency. Deviations occur that detract from performance and / or require verbal instructor prompting" (AETC Syllabus P-V4A-G, 2011). The second performance statistic was daily maneuver grades. This statistic follows the same convention as checkride scores but has a much lower average, because students tend not to be Excellent, or even Fair on their first attempts at a maneuver. All these scores are used to rank order students, and this rank order is used to assign students their follow-on aircraft and graduation awards. It is important to differentiate between the students for this reason, but improving their scores would indicate a higher quality of pilot production.

The counter data is also taken from TIMS, which records each different instructor with which a student flew, including both flight and simulator instructors. Under the 48 FTS Operating Instructions, a supervisor that flies with a student only once does not count against the student's continuity, but for this research it makes more sense to count each instructor that a student flies with as a counter. The fact that this data is even collected and displayed by TIMS is a good indication of the emphasis on this number in the past. This data was then adjusted because each time a student fails a checkride they fly with a flight commander from another flight for their "progress check" (who would be recorded as an additional counter). Binary data was collected on each student who failed a checkride, and then subtracted from the appropriate counter data to make the counter data an accurate picture of how many instructors a student flew with before their checkride.

The method of using checkride results to compare results in different methods of flight training is similar to a 2007 University of North Dakota research paper. This paper compared two different forms of training against the outcomes of the student's "stage checks," their form of a checkride. The paper also looked at the amount of repeat flights for each student, which is very similar to how this paper will look at checkride failure data (Schumacher, 2007). These similarities reinforce the idea that the methodology of looking at student's checkride scores and checkride failures is sound.

Other statistics were collected in an attempt to help control the data and check for other trends. Academics scores are based on the nine academic tests that students take, mostly before they start the flying portion of the syllabus. Flight hours are the amount of hours that each student spent at the controls of the airplane. In the T-1 there are normally two students on board, and one will be flying while the other is waiting in the back of the aircraft for their turn to fly. Class size represents the number of students in each class. Rank is the military rank of each student based on their pay grade. Second Lieutenants are represented by a one, First Lieutenants by a two and Captains by a three. Estimated date of completion shows when each class was scheduled to graduate. A syllabus dummy was collected because two different syllabi were effective during this time period. Flight Commander ranking shows where each student was ranked by their flight commander. Class rank shows where the student finished overall in the class after taking the appropriate percentages of each performance indicator.

A recent change in scheduling practices at Columbus AFB is the reason that this dataset could be particularly useful. Scheduling at the 48th FTS at Columbus prior to January 2012 was accomplished on the flight level (there are 4 flights in this particular squadron). In January, the responsibility for scheduling went to a squadron scheduling shop. This reduced the number of required schedulers from around eight to three. Schedulers went from scheduling around 10-20 events per day to scheduling around 60. This made a difficult part-time job which took up around 10 hours a week, into a difficult full-time job that took more like 40-60 hours per week.

This policy change also made it significantly harder for any scheduler to keep track of individual students. Under the previous system each scheduler would have 30-40 students to schedule, so it was feasible to know each one, whereas under the new system each scheduler was scheduling around 135 students. Another factor is that this squadron scheduler increased from around 12 possible instructors to around 65. The consequence was that each student was exposed to an increasing number of different instructors during their training. The following table shows the class averages of how many instructors (counters) with which each student flew. One can see that as the class number increases, the number of counters also increase, for the most part.

Table 1: Average Counters per Class

Class number	Mean Counters	Std Dev
1	31.92	2.20
2	34.00	2.99
3	33.85	3.09
4	30.77	3.19
5	33.20	1.74
6	30.42	2.62
7	32.38	2.53
8	35.75	2.24
9	35.35	3.04
10	33.50	2.35

Observations: 144

When we regress these statistics using the class numbers (1 through 10) as the independent variable and the counters as the dependent variable, we see a significant positive change in the transition phase (.176 coefficient and p-value of .002), an insignificant coefficient in the navigation phase (P-statistic of .341), and significant positive change in the overall counters (.232 coefficient and p-value of .008). This shows that overall numbers of counters did change over time such that from the first class to the fifth class there was an increase in the average number of instructors of one instructor, so over the whole dataset there was an average increase of two instructors. This means there is increased variability over this time period. This change was implemented at a certain point in time so it affected each class at a different point in their training. For this reason, linear regression or just comparing the means give us two fairly clear pictures on how the squadron was affected.

Other Questions

Another change over the time period of this study was a change in the syllabus. The SUPT has a separate syllabus for each phase of pilot training, and the syllabi are followed as closely as any other Air Force Regulation. Changes between syllabi are usually to address changes in how the Air Force trains pilots. One example would be an increase in the fidelity of the aircraft simulator would allow more events to be performed in the simulator, which would cut costs in jet fuel and aircraft maintenance. Syllabus changes happen on a discrete class-by-class basis. In this case the syllabus was changed between classes three and four. The classes under the old syllabus continued under the old syllabus and the new classes started with the new syllabus, so it can be studied using a dummy variable for each student, with 1 representing the new syllabus and 0 representing the old syllabus. Interesting changes in the new syllabus included: a decrease in amount of night time required, changes in checkride grading, and a change to the definition of the grade of Excellent. We will look at these changes to see how they will affect this research, and to see how small changes in wording affected different grades.

First the required night time was reduced from 4.0 hours to 3.5 hours (AETC Syllabus P-V4A-G, 2011). Although total hours were not reduced, this reduction in night time should have driven average hours down, but in reality there was an increase in average hours from the old syllabus to the new one. The increase was by approximately one flight hour, and was not found statistically significant using a dummy regression (p value of .112), but the difference in the means is still fairly interesting. The likely explanation is that the change did not reduce programmed syllabus hours, so ways were found to fly less night hours while still flying the same amount of total time.

Table 2: Average Flight Hours by Syllabus

Syllabus	Mean Flight Hrs.	Observations
Previous (0)	85.17	55
Current (1)	86.27	89

The second element related to the change in the syllabus is a change in grading and checkride grading criteria. The largest change was a change in the definition of the grade of "Excellent." Each grade has an explanation to help standardize grading. Excellent was previously described as "The student performs the operation, maneuver or task correctly, efficiently, and skillfully. Minor deviations occur that do not detract from overall performance" (AETC Syllabus P-V4-G(SGTO), 2010). The new definition deleted the portion stating that "minor deviations occur that do not detract from overall performance" (AETC Syllabus P-V4A-G, 2011). In flight training, where students attempt to hold a specific speed and altitude for long periods of time or during highly dynamic maneuvers, this is a sizeable change in the definition.

The other change was to checkride grading. Previously, all students started a checkride maneuver with a grade of excellent and the grade went down as they made mistakes. A recent inspection found this practice to be at odds with the syllabus. To fix this situation, check pilots started grading more in line with the course training standards, and since most students perform slightly above MIF (the required grade for proficiency) by the time they reach their checkride, they often score one grade above MIF. On the transition check this change reduced each the average student's score by eight points (out of an average of 540). To see if this was the case, we can compare checkride scores using the syllabus dummy variable to see if there was a change. We will also do this with daily maneuver scores to see if instructors are applying the new definition of Excellent. Results comparing the averages are displayed in Table 3.

Table 3: Summary Statistics Based on Syllabus Change

	Previous Syllabus (0)	Current Syllabus (1)
Trans Check Scores	0.9724	0.9135
Std Dev	0.0132	0.0190
Nav Check Scores	0.9737	0.9515
Std Dev	0.0153	0.0367
Daily Ride Scores	0.7041	0.6562
Std Dev	0.0247	0.0211
	Obs: 55	Obs: 89

As can be seen from these averages, evaluators are applying the new standards to the checkrides. The change in the definition can be seen by itself in the navigation check scores because the MIFs for the navigation checkride are all to the Good level. For the navigation check the average score falls around two percentage points. For the Transition check we see the change in the definition in excellent and the change to checkride grading procedures. The magnitude in this case is around a six percentage point drop.

These numbers definitely appear significant, but it will be useful to check significance using regression as well.

When we run these regressions using the syllabus dummy as the independent variable, we see that in each case the regression is statistically significant, although the navigation check has a very low R squared value. This is okay because we are not trying to explain everything about these variables, but only some of the variability. All of the coefficients are negative which shows that the new syllabus, represented by 1, lowered scores. We also see that the transition check is the most affected statistic, as it should have been, since the transition check was affected by both changes in grading and the change to the definition of excellent.

Table 4: Regressions Based on Syllabus Change

	Constant	Syllabus Coeff.	Std. Error	T-Stat	R Squared
Trans Check	0.972	-0.0588	0.0029	20.1	0.738
Nav Check	0.973	-0.0222	0.0052	4.26	0.107
Daily Raw Score	0.704	-0.0478	0.0038	12.37	0.515

Observations: 144

Since we have the data, it is also worth asking how class size affected the different student's checkride scores. This question has been asked in educational research many times over (Angrist, 1997), although these other studies involved a traditional classroom setting. Students in SUPT almost always have the same ratio of instructors-to-students during the flights, because almost all flights are conducted with two students and one instructor. Some students randomly receive a "singleton" flight where there is one instructor to one student, and almost all students fly two "singleton" flights before their navigation checkride, but for the most part the ratio during the flying portion stay at two to one. What we will see is whether the flight room dynamic of a larger or smaller class affects student's checkride scores. The classes in this sample ranged from 13 to 18 students, so there should be enough variability. We will also use the syllabus dummy and a dummy for the navigation checkride because those scores tend to be higher than the transition checkride.

Table 5: Effects of Class Size on Checkride Scores

	Coefficient	Std. Err.	T-Stat
Constant	0.9659	0.0168	57.31
Class Size	-0.0004	0.0012	0.29
Syllabus	-0.0400	0.0037	10.86
Nav Check	0.0239	0.0031	7.76

Observations: 288

According to Table 5 we see that class size does not have a statistically significant effect on checkride scores. The other two dummy variables are significant just like in many other regressions throughout this paper. The lack of significance on the class size coefficient indicate that these classes do not affect checkride scores, but that would not necessarily indicate that a class of 22 would be okay. It could also be that there was not enough variability in the class size variable, even though there is a 40% change when going from a 13 person class to an 18 person class.

Another question that we will examine is a long held theory that attempting more events will improve a student's scores. There are multiple angles behind this theory. First off, a student who flies more events thus gets more practice and should do better on checkrides. Since most students fly the same amount of flights, accomplishing more items per flight requires extra planning on the part of the student, and this could be an indicator of the level of effort or ability. Another theory is that instructors are more likely to grade events as Good, so more events graded as good should increase average daily maneuver scores. The counter-argument to this is that even if the student plans more events than average, they still have to perform well on each of those events. The points attempted statistic indicates a certain amount of points for every maneuver attempted throughout the program, so the average number is in the tens of thousands. A student who attempts one extra maneuver per flight would have a points attempted statistics approximately 220 points higher than their average counterpart. We will regress checkrides and daily flying scores against total points attempted to look at this relationship. Most instructors do not believe that these two are related, and even get frustrated because students will plan flights that last longer than they are supposed to in an attempt to maximize their points, but the results are rather interesting.

Table 6: Scores Regressed on Points Attempted

	Constant	Coefficient	Std. Error	T-Stat
Daily Scores	0.228	0.0000216	0.00000164	13.13
Checkride Scores	0.609	0.000016	0.00000139	11.81

Observations: 144

As we can see from Table 6, the value of points attempted in both instances is quite significant. The coefficients are quite small, but this is because the average number of points attempted is in the tens of thousands, while checkride scores are between 0 and 1. The R squared values are also quite high, showing that points attempted does explain a fair amount of the variability in the performance scores. This may be somewhat of a selffulfilling prophecy, because better students are usually able to accomplish more items, but this is significant for students trying to decide whether to "limit their exposed" or "maximize their points." Since both checkride scores and daily scores are significant we can postulate that the extra experience of accomplishing more events helps on checkrides, and that there may be something to the theory that attempting more events brings up your average because of how instructors grade.

Unused Variables

Before going into the model and the regressions we should talk a little more about some interesting variables that were left out of these regressions because of insignificance. Two variables that were collected but not used were the active duty and rank variables. Higher rank corresponds with higher maturity, but not necessarily with higher flying skills. These students would likely have higher flight commander rankings, but their flying scores could be anywhere. The active duty versus reservist or national guard student dummy was also insignificant. This probably stems from multiple reasons; the active duty students are mostly straight out of college with little prior flight experience, although some did study aviation in college or independently. The guardsmen and reservists are usually either experienced pilots or prior-enlisted officers. Those who are already pilots do quite well, like the active duty students with prior experience, but the others perform much like the active duty student group. The guardsmen and reservists groups also have a different set of incentives. Active duty students have to compete with each other because they are assigned to aircraft based on merit order, which means the first student gets their first choice, then the second student gets their first choice unless it was already taken, etc. (AETCI 36-2205v4, 2012). Guardsmen and reservists are already selected to fly a certain aircraft with a certain unit, so their incentives are to finish the program, and maybe win some graduation awards. Regardless of the reason, the dummy for guardsmen and reservist do no show up as significant in any of the regressions and thus are left out of the final research.

Another variable that was not used was flight hours. This was statistically significant, but likely for the wrong reason. Students who fail a checkride have to fly another flight, so poor checkride performance would correspond with higher flight hours. This is a statistics that would be worth studying by itself, but it did not help with the main idea of this particular research. Further research could collect data on which students fly extra flights or flight time during the daily flights and how their perform on checkrides differs, but this data was not collected.

Chapter III

METHODS AND RESULTS

To study the effects of counters on student performance we will first use OLS regression on the two checkride scores and daily maneuver scores to see if we get statistical significance for the counter coefficient. The entire sample will be used in these two regressions and will look like the following:

$$y = \beta_1 x_i + \beta_2 Syl + \varepsilon_i$$

Where *y* represents student performance on daily flights or a specific checkride, β_1 is the counter coefficient we are after, *x* represents the number counters of student *i,* and Epsilon represents the error term of student *i*. A syllabus term and a corresponding coefficient are included in this regression, because as was already shown, the syllabus change had significant effects on both daily scores and checkride scores. In all three regressions the syllabus coefficient was also shown to be significant with p-values less than 0.01, and so it will be included in all regressions including students in both syllabi.

The results of these OLS regressions are displayed in Table 7, and show that none of the regressions are significant on a 95% confidence level. To look a little further into these regressions we see that even though they are not significant, two out of three coefficient signs are negative. The magnitude of the coefficient is small, and if it were significant it would indicate, in the case of the navigation check (the most close to significant) that for each different instructor, the student's checkride score would fall by .0013, which is .04 of a standard deviation in the case of navigation checkride scores (the standard deviation is 0.032187) .

Table 7: OLS Regression

	Constant	Coefficient	Std. Error	T-Stat	p-value
Transition	0.964	0.000572	.00072	0.79	0.431
Navigation	0.998	-0.001360	.00126	1.09	0.278
Daily	0.720	-0.000480	.00062	0.79	0.434

Observations: 144

To look at the numbers in a slightly different way we will now run these regressions on all ten classes individually. The reason to run these all seperately is because each class goes through the course with a different core of instructors, at different times with different seasonal weather and many other changing conditions. These regressions will isolate the experiences that each class had and hold those constant. Although these regressions will be in smaller data pools, it is still worth looking to see if the trend exsists in each class.

Table 8: OLS by Checkride and Class

Class	Constant	Coefficient	Std. Error	T-stat	p-value	Obs
Transition						
1	0.979	-0.00097	.0031	0.32	0.755	14
2	1.005	-0.0019	.0011	1.82	0.093	14
3	0.970	0.0005	.0012	0.43	0.672	14
4	0.965	0.0002	.0022	0.13	0.901	13
5	0.895	0.0017	.0018	0.99	0.342	15
6	0.970	-0.0043	.0026	1.66	0.122	14
7	0.866	0.003	.0066	0.45	0.660	13
8	0.843	0.0048	.0027	1.75	0.101	16
9	0.895	0.00098	.0026	0.37	0.715	18
10	0.935	-0.0023	.0038	0.61	0.555	15
Navigation						
1	0.983	-0.0006	.0031	0.34	0.742	14
2	1.120	-0.0078	.0031	2.34	0.026	14
3	1.055	-0.0038	.0019	1.98	0.071	14
4	0.987	-0.0008	.0015	0.48	0.641	13
5	0.925	0.0019	.0039	0.49	0.629	15
6	0.955	-0.0001	.0069	0.01	0.989	14
7	1.020	-0.0036	.0019	1.89	0.420	13
8	0.819	0.0068	.0082	0.83	0.420	16
9	0.980	-0.0017	.0042	0.43	0.673	18
10	1.020	-0.0043	.0047	0.94	0.367	15

The OLS regression by class and by checkride is displayed in Table 8 and yields some interesting results. Only one class's counter coefficient (class 2) was found to be statistically significant, although three classes were very close to statistical significance at the 95% confidence level. The signs on the coefficents are a little more interesting. It was hypothesized that they would be negative for the most part because when we run the larger regression we notice that most of the signs are negative. Instead we find that around half of the signs are positive. The regressions that are statistically significant, almost significant, or have the higher R squared values do have negative coefficents, but it is still interesting that so many of the coefficients are positive. Due to the almost complete lack of statistical significance throughout these last regressions it would be better to use the regressions which have more data points to make any ruling on the effect of more instructors per student.

This series of regressions on a class level is a good sign for the fairness of the pilot training process. Since students are compared against their immediate classmates, and only one regression out of twenty was found significant, it appears that different number of instructors did not affect the classes merit orders.

Now that this data has been checked on a smaller level, we will increase the size of our dataset by combining results from both checkrides into one large column and bringing the corresponding counter data into another large column. We will continue to use the syllabus dummy, but we will also add a dummy variable to indicate the navigation check. This is necessary because as shown earlier, navigation check scores are higher on average than transition check scores. If this dummy was omitted then we would not be accounting for the fact that the navigation phase checkride scores should be higher according to the syllabus. This will double the size of our dataset and add a larger amount of variability.

Table 9: Regression of Combined Sample

	Coefficient	Std. Error	T-Stat
Constant	0.9883	-	-
Counters	-0.0019	0.0007	2.47
Syllabus	-0.0397	0.0032	12.55
Nav check	0.0306	0.0040	7.53

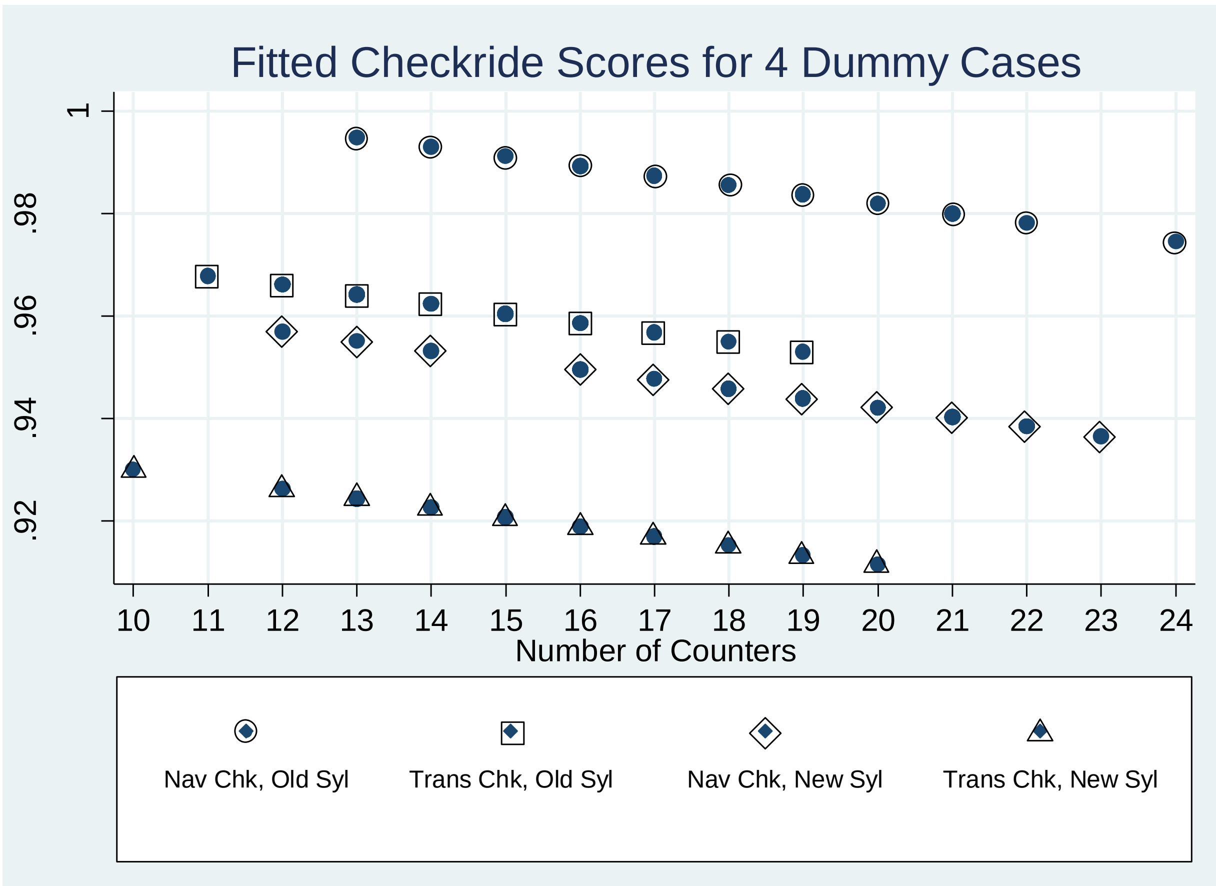
Observations: 288

This regression pegs all three variables as significant. The count variable is the least significant, but its significance is quite interesting at this point. Our data from earlier regressions was confirmed since the new syllabus variable corresponded to a drop in scores, and the navigation check indicated higher scores. The coefficents on these variables back this up with their signs, and the magnitudes of their constants are quite similar to the changes we noticed previously in Table 3. The count variable's coefficient is -0.00185, indicating that each new instructor that a student flies with may correspond to a drop in their checkride score of .00185.

This effect seems quite small, but when vying for position in a small class, every little advantage counts. If we take this a little further and figure out what the effect is on the average checkride we find that on the average checkride 540 points are attempted, and taking the 0.185% of that average total score equals a .999 point downgrade. The value .00185 also corresponds to about a tenth of a standard deviation for any of the checkrides in Table 3. Based on this information, the effect of the number of instructors is small, but significant.

To get a better picture of this we will look at a scatter plot of the fitted values based on the actual counter values. This means that the scatterplot only covers the portion of the data set for which we have data. In Figure 1 there are four distinct lines of scatter dots corresponding with the four different dummy states. The different lines are easily identifiable based on the previously proven information that the Navigation (Nav) Check has higher scores than the Transition (Trans) Check, and that the Transition Phase has a lower number of flights per phase, which means a lower number of counters. The prediction could be extended further, but based on this view the effects of extra instructors is apparent. The dot lines have negative slopes as expected based on the regression, but we can also see the interaction between the counter changes and the syllabus changes.

Figure 1: Fitted Scatter of Checkride Scores for the 4 Dummy States



Probit Regression

Overall scoring in pilot training is important, but another statistic that we are concerned with is checkride failures. There is an interesting relationship here because a student can fail a checkride and receive a better score than another student who passed their checkride. The difference is in the point system for checkrides and how unsatisfactory checkride performance is handled. Checkrides have a certain number of graded items, for example there are 52 items on the transition checkride. A student may be downgraded from Excellent to Good or from Excellent down to Fair. This would indicate that they scored either four out of five or three out of five points, respectively, on most items. If a Good is the required proficiency level, and the student receives a Good, then they get four out of five points. If they receive a Fair on that same maneuver then they fail the checkride and have to redo the event, but they earn three points (AETC Syllabus P-V4A-G, 2011). As mentioned previously, this requires another flight, and more jet fuel. The catch is that a student who gets downgraded from Excellent to Good on ten items would get a score of 40 out of 50 on those items and pass the checkride. A student who receives only one downgrade, but is downgraded to Unsatisfactory would receive a grade of 46 out of 50 on those same ten items, but would have to redo the checkride. To further test our hypothesis about instructor continuity we will check the data on checkride failures against the counter data.

For this regression we will run a probit analysis on the number of counters and a dummy variables indicating a checkride failure. A probit regression will be used because in this case the dependent variable is dichotomous, so the probit model makes more sense. The probit regression is specifically designed for use with a binary dependent variable and is non- linear. It uses the cumulative normal distribution and the output's coefficients are probabilities rather than slopes on a regression line. There is a pseudo R squared value along with the regression but it is much different than an R squared with normal regression and thus can not be compared side by side.

Using a probit regression to look at an aspect of pilot training was previous done by Reinhart (1998). Reinhard used the method along with OLS to look at certain biographical and selection criteria against training successes and failures of US Naval Academy graduates in pilot training. He used a two-stage Heckman regression to guard against selectivity bias, because there are not flight test scores for graduates who did not go to flight school. This should not be an issue for this research because we do have data for checkride failures, and because our sample is composed of pilots.

Table 10: Probit Regression on Checkride Failure Data

	Coefficient	Std. Error	z-score
Constant	-3.714	-	-
Counters	0.1716	0.0506	3.39
Navcheck	-0.6242	0.2633	2.37

Observations: 288

Looking at this regression we notice that the variable in question is significant. The z-stat is 3.39, giving a p-value of .001, and the coefficient on the counter variable is 0.171. In this case the fact that the coefficient is positive indicates that we are more likely to get a 1 for the dependent variable, which equates to a failed checkride. A little math is required to tease the magnitudes out of these coefficients since they are probabilities. We will show the change in probability of failing a checkride if a student has 12 different instructors and then if they have 16 instructors to illustrate the point.

$$prob_{12} = -3.714497 + .171687(12) = -1.1654253$$

$$converted\ to\ std\ norm = .12192344 = 12.19\%$$

$$prob_{16} = -3.714497 + .171687(16) = -.9757978$$

$$converted\ to\ std\ norm = .16458234 = 16.45\%$$

As we can see an additional four instructors raises a student's chances of failing a checckride by 4.46% for these specific numbers of instructors. Since this is not a linear model, the change in percentages would be different for different numbers of counters. This is a fairly small percentage, but it is statistically significant. Having been involved in the scheduling process, I would say that the gain in flexibility of being able to schedule students with different instructors outweighs the decrease in student quality. This is a good point to keep in mind though, and all other things being equal, limiting a student to a smaller number of instructors would be the better course of action according to these numbers.

Difference-in-Difference

The last method we will use to see if there is an appreciable effect of counters on student pilots is difference-in-differencing. We are lucky in that we have two different data points for each student, the transistion and navigation checkrides. If we difference the number of counters in each phase and then also difference the checkride scores we should eliminate personal traits of each student and be able to more accurately isolate the effects of the number of counters. The issue with this plan is that the phases of training are different. Transiton has only 12 lessons, while navigation has 16. There are also grading differences in the two checkrides, if we look at the averages. To fix these issues we will demean the data, convert everything into standard deviations and then difference them by subtracting navigation check scores from transition check scores and navigation block counters from transition block counters. This way we will see if a change in the amount of counters per block affects the difference in the scores.

Table 11: Difference in Differencing Results

Syllabus	Constant	Coefficient	Std Error	T-Stat	p-value	Obs
0	-0.0501	-0.1809	0.1242	1.46	0.151	55
1	-0.0134	0.0865	0.1083	0.8	0.427	89

Looking at the results here we do not see any statistical significance. Different regressions were run for the two different syllabi so that the changes in grading practices would not affect the results. We also see different signs on the two different regressions which reinforces the insignificance of these regressions. There are some differences between the two checkrides that could make this regression less effective. For example the two checkrides test somewhat different skills, and the expectation is much higher for the second checkride. They should be similar enough though because in both cases the students are taught what they should know for the checkride.

CHAPTER IV

SUMMARY

In this research paper we have examined the effects of multiple instructors on student's checkride scores in Phase III of United States Air Force Pilot Training. Multiple regression, difference-in-differencing, and a probit model were used to compare results from 144 students over an approximately one-year period. Daily maneuver scores and two different checkride scores were collected along with the corresponding number of instructors. Other data was collected, but the only data found to be useful and statistically significant were dummy variables indicating changes in the syllabus and the two different checkrides.

We checked a few hypotheses on the effects of a syllabus change including a change in night flying requirements, changes in checkride grading, and all grading. We found that scores fell from one syllabus to the next because of a change in the definition of the grade of Excellent. Changes in checkride grading procedures lowered the average scores of the transition checkride also. When we checked if a drop in night hours required affected total time per student we found that it did not decrease total time, and in fact that total time increased from the old to the new syllabus. Using this information the syllabus dummy variable was used when running further regressions so as not to omit an important variable. A dummy was also used indicating the navigation check when both checkrides were run together.

It was next important to show that as time passed the number of counters increased per student. We looked at averages and ran a regression to show that this was the case. This helped increase the variability of the dataset, and showed what the effects were of changing the squadron from flight scheduling to centralized scheduling. This settles a contested question in the squadron about the effects of squadron scheduling. Whether for better or worse, we can say for certain that students flew with more instructors after the changeover to centralized scheduling.

We then started on the main question: "what is the effect of higher numbers of instructors per student?" We used checkride scores and binary data on checkride failures. When the checkrides were regressed separately and by class the data indicated that the relationship between the number of instructors and scores was statistically insignificant. When the two checkrides were combined and put in the same regression it was statistically significant. This regression was controlled by a dummy for the navigation checkride, because navigation checkride scores are normally higher than transition check scores. This regression had a negative coefficient on the counter variable and was modeled based on the four different options for the two dummy variables. The modeling showed a downward trend in three cases, but a positive trend in the forth. Given these regressions there is evidence that checkride scores are related to the number of instructors with which a student flew. When we look at the largest regression we see that each additional instructor equates to around a one point deduction in the checkride score. Although the data is somewhat mixed on this, the larger sample supports the idea that more instructors equal lower checkride scores.

A probit regression was then untilized to test whether more instructors increse the probability of a checkride failure. The results of this were strongly positive that more instructors increase likelihook of a checkride failure. This is also a more useful measurement because checkride failures almost always lead to another flight, whereas checkride scores do not. We found that in the normal range of counters, one counter increased chances of a checkride failure by approximately one percentage point.

The difference in differencing method did not show statistical significance, but this may be due to differeces of the two different checkrides and phases of training. The data was converted into standard deviations to get around some of those difference, but the regression was still insignificant.

Overall, an incease in counters does affect student performance in the ranges tested. The effect is small, but statistically significant. The application for this research is that flight training units or schools should attempt to limit students exposure to different instructors. The affects are small enough, and the tradeoff are such that changes to the syllabus to limit students to less instructors should not be required. If reducing the number of counters is not possible, schedulers and flight commanders should realize the effects they incur each time they assign a student to fly with a new instructor.

REFERENCES

AETCI36-2205 *Formal Flying Training Administration and Management – T1A, T6A, T38C, and TH1H*, Volume 4, Randolph AFB, TX:HQ AETC/A3FP, 24 July 2012.

AETC Syllabus P-V4-G(SGTO), *T-1 Joint Specialized Undergraduate Pilot Training*, Babcock, Steve, POC, Cert. Charles E. Catoe, ed. John P. Fosdick., Randolph AFB, TX: HQ AETC/A3FP October 2010.

AETC Syllabus P-V4A-G, *T-1 Joint Specialized Undergraduate Pilot Training*, Babcock, Steve, POC, cert. Charles E. Catoe, John R. Fosdick, Randolph AFB, TX: HQ AETC/A3FP, June 2011.

Angrist, Joshua D. and Victor Lavy, *Using Maimonides Rule to Estimate the Effects of Class Size on Scholastic Achievement*, NBER Working Papers Series, National Bureau of Economic Research, January 1997.

Blower, D.J., *A Cost-Benefit Analysis of the Impact of Selection Testing on Advanced Flight Training*, Naval Aerospace Medical Research Laboratory, Pensacola, FL, 1997.

Christensen, Cody J. and Bryant H. Dunn, *Fleet Characteristics of Collegiate Aviation Flight Programs*, Collegiate Aviation Review, Vol. 29, Issue 2, (2011).

Columbus Air Force Base Factsheet, 14 Flying Training Wing Public Affairs Office, http://www.columbus.af.mil/library/factsheets/factsheet.asp?id=12792, Posted 1 February 2013 (accessed on 5 February 2013).

Columbus Air Force Base History Factsheet, 14 Flying Training Wing Public Affairs Office, http://www.columbus.af.mil/library/factsheets/factsheet.asp?id=12792, Posted 5 October 2006 (accessed on 2 February 2013).

Fritzinger, Thomas J., *Economies of Scale in Air Force Undergraduate Pilot Training*, Thesis for Air Force Institute of Technology, Sept 1980, p. 5. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA093199&Location=U2&doc=GetTRDoc.pdf (accessed on November 5, 2012).

Gardlin, Gene R. and Thomas E. Sitterley, *Degradation of Learned Skills, Manned Spacecraft Center*, NASA, Seattle, WA, June 1972.

Government Accountability Office, *Aircraft Acquisition: Affordability of DOD's Investment Strategy*, Washington DC, NSIAD-97-88, Sep 8, 1997.

Hall, David J., *A Case for Teacher Continuity in Inner-City Schools*, The School Review, Vol. 80, No. 1, (Nov. 1971) University of Chicago Press, Chicago, p. 27-47.

Jackson-Evers International Airport, Airport Information Database, Airnav.com, http://www.airnav.com/airport/KJAN (Accessed on 28 Nov 2012).

Jehn, Christopher, *Congressional Budget Office Testimony, Pilot Retention: Issues and Possible Solutions*, Statement before the Subcommittee on Military Personnel, Committee on Armed Service, US House of Representatives, 4 March 1999.

Lynch, William E. Capt., USAF, *A Meta-Analysis of Pilot Selection Tests: Success and Performance in Pilot Training*, Defense Technical Information Center, 2 Mar 1992. http://www.dtic.mil/dtic/tr/fulltext/u2/a246623.pdf (accessed on November 1, 2012).

Martz, Lt. Joshua, 48 FTS Scheduler, Email message to author, 30 November 2012.

Reinhart, Paul M., *Determinants of Flight Training Performance: Naval Academy Classes of 1995 and 1996*, Naval Postgraduate School, Monterey, CA, August 1998.

Rokicki, Stephen M., *Undergraduate Pilot Training Duty Schedules and Aircrew Fatigue*, USAF School of Aerospace Medicine, Brooks AFB, TX, July 1981.

Savion-Lemieux, Tal and Virginia B. Penhume, *The Effects of Practice and Delay on Motor Skills Learning and Retention*, Department of Psychology, Laboratory for Motor Learning, Cognitive Learning, and Neural Plasticity, Concordia University, Springer-Verlag, 1994.

Schumacher, Peter M. and Lease, Jared, *Evaluating the Effectiveness of Scenario Based Training in the Collegiate Flight Training Environment*, Joint Training Standards Development for New Technology General Aviation Aircraft Task 4, FY 2004 FITS Grant, Second Follow-on, February 15, 2007, www.faa.gov/**training**_testing/**training**/fits/**research**/.../SBT_final.pdf (accessed Nov 20, 2012).

Vance Air Force Base Factsheet, 71st Flying Training Wing Public Affairs Office, http://www.vance.af.mil/shared/media/document/afd-090227-127.doc, Current as of March 2011 (accessed on 5 February 2013).

Wilson, Wayne Bruce, *The Effect of Prolonged Non-flying Periods on Pilot Skill in Performance of a Simulated Carrier Landing Task*, Naval Postgraduate School, Monterey, California, September 1973.