A META-ANALYSIS OF PILOT SELECTION TESTS: SUCCESS AND PERFORMANCE IN PILOT TRAINING

THESIS

William E Lynch, B.S., M.L.A.
Captain, USAF

AFIT/GLM/LSM/91S-44

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

1

AFIT/GLM/LSM/91S-44

A META-ANALYSIS OF PILOT SELECTION TESTS: SUCCESS AND PERFORMANCE IN PILOT TRAINING

THESIS

William E Lynch, B.S., M.L.A.
Captain, USAF

AFIT/GLM/LSM/91S-44

The views expressed in this thesis are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

AFIT/GLM/LSM/91S-44

A META-ANALYSIS OF PILOT SELECTION TESTS: SUCCESS AND PERFORMANCE IN PILOT TRAINING

THESIS

Presented to the Faculty of the School of Systems and Logistics

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Logistics Management

William E Lynch, B.S., M.L.A.

September 1991

Preface

The purpose of this research was to systematically evaluate the literature available on predicting the success of undergraduate military pilot trainees.

The literature was categorized by predictors to determine the number of each type of predictor. Once categorized, it was found that there were more studies on cognitive predictors for the Navy and the Air Force than any other type of predictor.

A meta-analysis was performed on the Air Force Officer Qualifying Test Pilot Composite and the Navy/Marine Aviation Selection Battery Flight Aptitude Rating, with respect to their ability to predict the successful completion of pilot training.

In performing the research and writing for this thesis I received a great deal of help from others. Dr. Dennis Campbell, my thesis advisor helped me through some frustrating phases of the process. I would also like to thank Dr. Guy Shane, whose assistance with both the research and write-up was critical in completing this project.

Finally, to the two most important people in my life, my wife Gina and son William Edward, whose love, patience, and support kept me motivated through countless hours of frustration.

William E. Lynch

Table of Contents

List of Tables

Abstract

The purpose of this study was to determine if the characteristics measured by the Air Force Officer Qualifying Test Pilot Composite and Navy/Marine Flight Aptitude Rating were significantly correlated to the successful completion of flight training. Meta-analysis was used to: calculate a mean weighted average correlation, and correct for sampling error, error of measurement, restriction of range, and dichotomization. Over 200 studies were considered for the meta-analysis.

The results indicate that both the uncorrected and fully corrected weighted mean correlations for a group of nine Air Force studies were statistically significant ($p<.0001$). The partially corrected (sampling error and dichotomization) correlation for a group of eight Navy studies was also statistically significant ($p<.03$), while the uncorrected weighted mean correlation was not significant ($p>.05$). There was no significant difference between the magnitude of the correlations (uncorrected and corrected) between the Navy and Air Force groups.

The findings of this research indicate that both the Air Force Officer Qualifying Test Pilot Composite and Navy/Marine Flight Aptitude Rating are useful in selecting those candidates who are more likely to complete pilot training.

A META-ANALYSIS OF PILOT SELECTION TESTS: SUCCESS AND PERFORMANCE IN PILOT TRAINING

I. Introduction

Introduction to the Chapter

This chapter serves the purpose of introducing the research problem. It contains an introduction to the problem, the research hypothesis, research questions, scope and limitations of the research, assumptions made by the researcher, definitions of key terms, and a summary.

Introduction to the Problem

In recent years we have seen an increasing emphasis put on decreasing the money spent by the military. The 1992 National Defense budget reflects an increase of 2.6 billion dollars from 1991. With inflation, this budget represents a real decrease in military spending of approximately 12 billion dollars (Collender, 1991). This real decrease in military spending will make the efficient training of pilots even more critical than it is today.

A validated model for reliable pilot selection could be a major player in solving the problem of high attrition historically experienced in Undergraduate Pilot Training (UPT). It could also significantly reduce the associated costs mentioned below.

Attrition. The problem of pilot attrition is one that affects every military organization conducting flight training. For the United States Air Force, the problem has multiple effects. One effect is the cost of training (estimated to be between $65,000 to $80,000) a pilot candidate before elimination from the program (Siem et al., 1988). Another effect is failing to meet the manning requirements that determine the number of training positions programmed for UPT. If the attrition rate continues to be higher than expected, a shortfall of pilots to fill the manning requirements will occur.

If the predicted attrition rate is not accurate, the result could be a serious shortfall in operational manning, threatening the strength of our national defense. The shortfall would cause a strain on the flying squadrons, as well as support organizations. Reducing this attrition rate, with a valid and consistent selection strategy that accurately predicts the success of pilots in UPT, will decrease the cost of meeting operational manning requirements.

USAF Pilot Selection Process

Uncommon Selection Criteria. The Air Force recruits undergraduate pilot candidates from three different sources: Officer Training School (OTS), the Air Force Reserve Officer Training Corps (AFROTC), and the United States Air Force Academy (USAFA) (Davis, 1989:4). These sources do not use common selection criteria, thus increasing the error variance of the selection process across sources.

Officer Training School (OTS). Officer Training School takes place at Lackland AFB, Texas. During the 120 day training, OTS performs both officer training and flight screening. Acceptance as a candidate requires possession of a college degree, passing a medical examination, and attaining a qualifying score on the Air Force Officer Qualifying Test.

Those selected as pilot candidates take a "portabat" test (a computerized video device very similar in structure to an arcade video game) and go through flight screening before OTS. The portabat tests hand and eye coordination and the learning curve of the individual. The learning curve of the individual is tested by determining how well the subject can keep two bars (one vertical and one horizontal) within target parameters on the screen (Eisen, 1988:22).

The flight screening phase takes place at Hondo Field, Texas. Candidates undergo a 16 day program, completing nine hours of ground training and 14 hours of flying time. The T-41 aircraft is used to evaluate the candidate's flying skills. If the candidate receives a satisfactory rating (rated as either satisfactory or unsatisfactory) for flight screening phase, they begin OTS.

OTS then serves as an additional screening device, measuring the candidate's ability to operate in a stressful environment. Successful completion of OTS allows the candidate to enter Undergraduate Pilot Training (Eisen, 1988:23).

Air Force Reserve Officer Training Corps (AFROTC). AFROTC candidates must also successfully complete the Air Force Officers Qualifying Test(AFOQT) and a medical examination to compete for an undergraduate pilot training position. Their grade point average, Scholastic Aptitude Test scores, and unit commander ratings are also considered. A central selection board held at Maxwell AFB, Alabama, makes the final selections for ROTC pilot training.

In order to compete for undergraduate pilot training, positions, AFROTC applicants must attain a minimum percentile score of 25 on the pilot composite of the Air Force Officer Qualifying Test, and 10 on the navigator-technical composite. In addition, the combined score of both must total at least 50 (Arth et al., 1990:1). For example, a score of 10 on the Navigator-Technical composite would require that a candidate score at least 40 on the pilot composite to qualify. Upon selection for the flight training program, cadets enroll in the Professional Officer Course (POC). The POC begins in the candidate's junior year of college.

In addition to meeting the requirements above, candidates must also successfully complete flight screening. Light aircraft training for AFROTC follows the same format as Officer Training School. AFROTC candidates attend flight screening at either the USAF Officer Training School flight screening facility at Hondo, Texas, or Embry-Riddle Aeronautical University, Daytona Beach, Florida (Eisen, 1988:24). Successful completion of light aircraft training marks the end of the flight screening phase. After college graduation and commissioning, the candidate enters into the Undergraduate Pilot Training Program.

United States Air Force Academy (USAFA). The United States Air Force Academy is the third source of UPT candidates. Along with the normal requirements associated with applying to any undergraduate institution, the academy requires that the candidate complete a series of physical tests. Medically qualified candidates who, successfully complete the Pilot Indoctrination Program (consisting of approximately seven hours of "airmanship academics" and 20 hours of flying time), and receive a positive recommendation from flying supervisors, enter UPT after graduation (Eisen, 1988:25).

US Navy Pilot Selection Process

The Navy has two flying screening programs; the Naval Academy and the Aviation Officer's Candidate School. One major difference between the Navy's program and that of the Air Force is that the Navy does not require the candidate to have a college degree. The candidate may apply to enter the Naval Aviation Cadet program after completing two years of college. The training received by candidates without a degree is identical to those who possess a college degree. However, candidates who do not have a degree do not receive their commission until completion of both Aviation Officer's Candidate School and Naval Aviation Flight Training.

Research Problem

The problem analyzed by this study is that there is no single standardized method developed for pilot selection. Research completed on predicting the successful completion of undergraduate flight training, has shown conflicting results. For example, the uncorrected correlations used in the meta-analysis for the USAF studies ranged from .09 to .21.

Research Question

The research question addressed in this thesis is: Is there an identifiable correlation between the characteristics measured by the selection tests and the successful completion of pilot training (inc[illegible]g both the Air Force and the Navy)?

Subsidiary Research Questions.

1. Is there a measurable difference in mean correlations between results presented by the US Air Force and those presented by the Navy?

2. Will the meta-analysis procedure significantly ($p<.05$) increase the magnitude of the correlation between

the predictor (AFOQT Pilot Composite and Navy Flight Aptitude Rating) and criterion (Success in pilot training)?

Research Hypothesis

The null hypothesis asserts there is no significant relationship between the characteristics measured by selection tests and successful completion of undergraduate pilot training for both the Air Force and Navy samples. The test hypothesis for the Air Force sample is as follows:

Ho: The mean corrected correlation between the AFOQT Pilot Composite and success in undergraduate pilot training is not significantly ($p<.05$) different from zero (equals zero).

Ha: The mean corrected correlation between the AFOQT Pilot Composite and success in undergraduate level is significantly ($p<.05$) different than zero.

The test hypothesis for the Navy sample is as follows:

Ho: The mean corrected correlation between the Navy Flight Aptitude Rating and success in flight training is not significantly ($p<.05$) different from zero.

Ha: The mean corrected correlation between the Navy Flight Aptitude Rating and success in flight training is significantly ($p<.05$) different from zero.

Scope

This study relies on the results of previous research studies concerning Air Force and Navy pilot selection. The literature review involved the gathering, reviewing, and coding of studies for use in the meta-analysis procedure. The results are cumulated and summarized in chapter 4. The main characteristics of interest are the predictors measured by the various tests and their relationships to the attrition of undergraduate pilots.

Limitations

This research of studies in pilot selection is not all inclusive. It consists of meta-analyses of other research that may contain artifacts. Not all artifacts are corrected for by the meta-analysis techniques. Therefore, it is likely that the correlation derived through meta-analysis will be a conservative estimate.

Assumptions

A primary assumption of this study is that previously conducted studies used for the meta-analysis reflect an accurate transcription of the results in each experiment.

Another assumption of this research is that the correctable artifacts occurring in the previous studies (e.g. sampling error, error of measurement, restriction of range, etc.) can be demonstrated and corrected for through the meta-analysis procedure. This of course, would make the cumulation of the results (combining of correlations across studies) much more accurate. If correctable artifacts can be demonstrated, the variance caused by them will be accounted for, and the new correlations will be more accurate and meaningful. A more complete discussion of these procedures is presented in the next chapter.

Definitions of Key Terms

Artifacts are those flaws in the research design or inherent limitations in analysis procedures that cause the data to produce less-than-accurate results (i.e. sampling error, restriction of range, error of measurement, etc.). See Chapter 2 for further explanation of the various kinds of artifacts (Hunter and Schmidt, 1990:43).

Attrition, for the purposes of this research, refers to all instances in which undergraduate pilot candidates fail to complete undergraduate pilot training, for any of numerous possible reasons (e.g. self-initiated, medical, academic elimination, etc.).

Dichotomization refers to a variable being divided into two choices (i.e., success/failure). The magnitude of the correlation between the predictor and criterion of a study would be greater if a continuous scale were used. For example, each candidate is assigned a grade between zero and 100 based on flying ability, those who score above 70 continue their training. Chapter II develops this artifact further. (Hunter and Schmidt, 1990:46)

Error of Measurement is an artifact that comes from the degree to which the instrument contains random error. This is the unreliability of the measurement, or the degree to which the measurement does not give consistent results, when all other factors remain the same (Hunter and Schmidt, 1990:44).

Meta-analysis is a term coined by Glass (1976) that refers to "the quantitative cumulation and analysis of descriptive statistics across studies, without requiring access to original study data." (Hunter et al., 1980:137)

Reliability refers to "the degree to which a measurement is free of random or unstable error." (Emory, 1980:132) It is the consistency of the measurement.

Restriction in Range refers to a study sample (such as, all those who have been accepted for pilot training) that has been pre-selected and does not represent the overall population (all those who apply for pilot training). This is a commonly occurring artifact corrected for through meta-analysis techniques. For example, many of the pilot selection research results were derived by studying those who were already successful in getting accepted to undergraduate pilot training (Hunter *et al.*, 1982: 61).

Sampling error is the degree to which the sample falls short of representing the true characteristics of the population (Hunter et al, 1982:40-41).

Validity is "the extent to which differences found with a measuring tool reflect true differences among those being tested." (Emory, 1980:128) Validity refers to the test's ability to measure the desired characteristic.

Chapter Summary

This chapter describes the complexities, problems, and hypotheses associated with researching of the pilot selection process. It contains an introduction to the problem; including candidate attrition and varied commissioning sources. The chapter also covers the research hypothesis, research questions, scope and limitations of the research, assumptions made by the researcher, and definitions of key terms. The next chapter will cover the details of the methodology used to accomplish the research presented in this thesis.

II. Methodology

Introduction to the Chapter

This chapter details the plan for accomplishing the research. It begins with an explanation of the meta-analysis procedure, and includes important aspects of meta-analysis: cumulation procedures, study artifacts and their impact on study results, integration of research findings across studies, and measures required to complete the meta-analysis. The results of the literature search and selection of predictors to study are also addressed.

Meta-analysis

Meta-analysis is simply a statistical analysis of previous statistical analyses. It integrates statistics of prior studies to get a weighted best estimate of the correlation being studied. The purpose of doing meta-analysis is to improve the statistical power of a relationship between variables. Glass states:

> By recording the properties of studies and their findings in quantitative terms, the meta-analysis of research invites one who would integrate numerous and diverse findings to apply the full power of statistical methods to the task. Thus it is not a technique; rather it is a perspective that uses many techniques of measurement and statistical analysis. (Glass et al., 1981: 21)

Davis and Steel divide meta-analysis into three steps:

(1) Conducting an exhaustive search on the topic of the study;

(2) Extracting and coding the findings and characteristics of the studies; and,

(3) Cumulating and summarizing the findings using any number of known inferential and/or descriptive data analysis procedures (Davis and Steel, 1990: 2).

By combining the results of many research studies, it is possible to recognize a relationship that was not otherwise apparent. Davis and Steel state that the advantage of using meta-analysis is "that by comparing results across studies one avoids problems inherent in individual studies, e.g., inadequate sample size and problems with statistical power" (Davis and Steel, 1990: 3).

Cumulation Procedures. Hunter et al., categorize the cumulation of results across the studies into a five step process:

> (1) calculate the desired descriptive statistic for each study available, and average that statistic across studies;
> (2) calculate the variance of the statistic across studies;
> (3) correct the variance by subtracting the amount due to sampling error;
> (4) correct the mean and variance for study artifacts other than sampling error; and,
> (5) compare the corrected standard deviation to the mean to assess the size of the potential variation in results across studies in qualitative terms. If the mean is more than two standard deviations larger than 0, then it is reasonable to conclude that the relationship considered is always positive (1982:28).

Study Artifacts and Their Impact on Study Outcomes

Hunter and Schmidt (1990:44), identify 11 artifacts that alter the size of the study correlation in comparison with the actual correlation. They are: sampling error, error of measurement in the dependent variable, error of measurement in the independent variable, dichotomization of a continuous dependent variable, dichotomization of a continuous independent variable, range variation in the independent variable, range variation in the dependent variable, deviation from perfect construct validity in the independent variable, deviation from perfect construct validity in the dependent variable, reporting on transcriptional error, and variance due to extraneous factors (Hunter and Schmidt, 1990:45).

This research addresses the four major artifacts identified by Hunter and Schmidt as causing the largest variance: sampling error, error of measurement, dichotomization, and range restriction. Error of measurement, and range variation can be corrected with respect to the predictors in this study. Dichotomization is corrected for with respect to the dichotomous criteria (success/failure in pilot training). The following paragraphs describe each of these artifacts in greater detail.

Sampling Error. Emory describes a "good sample" as one whose design "represents the characteristics of the population it purports to represent" (Emory, 1980:148). How well the sample represents the population depends on both its accuracy and precision. The term "accuracy" represents the degree to which the sample is free from systematic error or bias. "Precision" represents the degree to which random error is absent in the sampling process. The degree of sampling error is inversely related to the degree of precision in the sample. The sampling error randomly appears on either side of the correlation coefficient. It is reasonable to conclude that net sampling error decreases as the sample size becomes larger (based on the fact that the random error on both sides of the correlation coefficient will tend to move toward the true mean: Central Tendency Theorem). Thus, a benefit of meta-analysis is the that as sample size increases, sampling error decreases (Hunter and Schmidt, 1990:44).

Error of Measurement. The error of measurement is an artifact that comes from the degree to which measures taken with the instrument contain random error. This is unreliability of the measurement (the degree to which the measurement does not give consistent results, when all other factors remain the same).

An error of measurement can occur in either the criterion, the predictor, or both. Hunter and Schmidt state that "simple error of measurement is the random measurement error assessed as unreliability of the measure" (Hunter and Schmidt, 1990:44,46). The actual correlation (the true correlation measured by a perfect study) between the predictor and criterion is equal to the observed correlation (with associated variance) divided by the square root of the reliability of the measurement:

$$r_c = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}} \qquad (2\text{-}1)$$

where r_{xy} is the correlation between the predictor and criterion, r_{xx} is the reliability of the predictor measurement, and r_{yy} is the reliability of the criterion measurement. In this study, r_{yy} is 1, since there is no reliability measure for the pass/fail criterion measure. This results in a conservative estimate for the corrected correlation. For example, if the reliability of the measure is .81, the observed correlation would equal .90 (square root of .81) multiplied by the actual correlation. This reduces the correlation by .10 through artifactual attenuation (Hunter and Schmidt, 1990:46).

If both the criterion and predictor measures have less than perfect reliability, the actual correlation equals the observed correlation divided by the square root of the reliability for the dependent variable multiplied by the square root of the reliability for the independent variable. For example, if the predictor measure had a reliability of .81, while the criterion had a reliability of .64, then the multiplicative effect would reduce the correlation to .72 (square root of .81 multiplied by the square root of .64). This would reduce the observed correlation by .28. With this multiplicative effect it is important to evaluate the reliability of the measurements (Hunter and Schmidt, 1990:45).

Dichotomization. If a continuous variable is dichotomized (divided into two categories, such as, "completed UPT" and "failed to complete UPT") by the researcher, then the given correlation for the dichotomous variable will be less than that of a continuous variable. The effect of using a dichotomous measure versus a continuous variable depends on where the continuous variable is split. The smallest reduction in correlation occurs when the continuous variable is split 50-50. This research involves a criterion that is dichotomous (pass/fail in undergraduate pilot training). According to Hunter and Schmidt, a 50-50 split reduces the correlation by .20 (Hunter and Schmidt, 1990:46-47).

Range Restriction. By restricting the range of the sample (for example, by testing only candidates already selected to attend UPT) the researcher is decreasing the magnitude of the correlation. Formulas used to correct for this problem are covered in the discussion and results portion of this thesis (Chapter 4).

All studies in this research involved a sample selected to attend undergraduate pilot training. Because the subjects had been pre-selected, and the sample is restricted in range, the observed correlation is misleading. The correlation demonstrated on a restricted population is smaller than the correlation of an unrestricted population.

The studies included in the meta-analysis only contained criterion scores for those s lected to attend pilot training.

Cumulating Correlations Across Studies

Through meta-analysis it is possible to correct for many of the sources of error that affect the correlation coefficient (e.g., sampling error, error of measurement, and range variation).

Sampling error is corrected by considering the sampling error for the meta-analysis as equal to the average of the sampling errors in each study. Simply put, if there are 50 studies with a total sample size of 5000, then the sampling error for the correlation is estimated as the calculated sampling error for a sample size of 5000 (Hunter et al, 1982:33).

It is also necessary to know the variance of the correlations across the studies caused by the sampling error. The effect of sampling error on the variance is to add a known constant -- sampling error variance. Once calculated, the error variance is subtracted from the observed variance to get an estimate for the variance of the population correlations. The objective of meta-analysis with regard to sampling error is to transform the distribution of observed correlations into a distribution of population correlations. "We would like to replace the mean and standard deviation of the observed sample correlations by the mean and standard deviation of population correlations" (Hunter _et al_., 1982:33-34).

Once the variance caused by sampling error is corrected, the real variance is transparent. This allows researchers to estimate the level of real variance across the studies. After variance caused by the sampling error, the error of measurement, and the range variation are corrected, it is possible to investigate moderating variables (those variables that would naturally affect the variance across studies) For example, there may be a difference caused by the size of the organization (Hunter et al, 1982:36).

Criteria for Selection of Studies

Certain types of data must be present to qualify a study for use in the meta-analysis procedure. The following criteria were required:

1. The study must present a conclusion that can be transformed into a common statistic (e.g., Pearson's r, biserial, point-biserial, etc.);

2. The sample size must be reported;

3. The study must contain a predictor variable (e.g., AFOQT Pilot Composite or Age of the candidate); and,

4. The study must contain a criterion variable (e.g., the successful completion of UPT or the final grade in a phase of training).

The failure to meet any of the four criteria described above excludes the study from further examination.

Pilot Selection Predictor Categories

A literature search yielded over 200 research studies that examined the relationship between a predictor variable and measure of successfully completing undergraduate pilot training. Of these 200 studies, a total of 79 met the required selection criteria and became candidates for the meta-analysis.

The predictors of the 79 studies were grouped into four categories: Demographic, Psychomotor, Personality, and Cognitive. In the discussion of each category that follows, it should be noted that some research studies investigated more than one predictor variable and qualify for inclusion in more than one category. Therefore, the number 79 will be exceeded if all four categories are summed.

Demographic. 17 studies included data on demographic predictors, such as, age, or gender. While some of the predictors demonstrate potential in selecting candidates more likely to complete pilot training, a meta-analysis would probably not add any insight on their statistical significance. For example, the age of the candidate was compared with the successful completion of pilot training in six studies. This was the largest number of occurrences of any demographic predictor. The fact that they were not used for the meta-analysis, by no means, detracts from their validity as predictors.

Psychomotor. 20 studies out of the 79 included psychomotor predictors. The psychomotor studies yielded only six different predictors, and the total number of studies that looked at any one predictor (with the same criterion) was five. Once again, the fact that a meta-analysis was not done on these studies does not reflect on the validity of the psychomotor predictors.

Personality. 37 studies out of the 79 included personality predictors. The personality studies yielded 24 different personality inventories (multiple subscales), with more than 150 different predictors. The problem here was the total number of studies that investigated any inventory was four. Many of the inventories measure similar personality traits (e.g., carefulness or risk-taking), but it was hard to determine the predictors were, in fact, the same. For example, two studies might report on a predictor that they both call "risk-taking", but then define it differently.

It was also apparent from statements in the literature that many inventories were developed using previous inventories. The problem was that these studies did not delineate how the new inventory related to the old.

Cognitive. Cognitive predictors yielded 23 studies that used some sort of cognitive testing (e.g., Air Force Officer Qualifying Test) to predict the success of undergraduate pilot candidates in pilot training. Within these 23 studies there were 17 that included either the AFOQT Pilot Composite or FAR as predictors of success, and had the same criterion (successful completion of pilot training). Of the 17 studies, 9 were Air Force studies and 8 were Navy studies.

It is important to note that the cognitive tests had high reliabilities, and were relatively easy to relate between the Air Force and Navy. This makes the meta-analysis easier to perform, because reliabilities can be used to correct for error of measurement. Another key factor in the cognitive studies is that the pilot composite of the Air Force Officer's Qualification Test (AFOQT) measures similar abilities as the Flight Aptitude Rating (FAR) of the Navy/Marine Corps Aviation Selection Battery. The number of useable studies, the high reliabilities of the measures, and ease of combining the studies resulted in the cognitive measures being chosen for the meta-analysis.

Chapter Summary

This chapter describes the methodology used to accomplish a meta-analysis of pilot selection studies. It contains an explanation of meta-analysis, cumulation procedures, study artifacts and their impact on study outcomes, the integration of research findings across studies, measures required to complete the meta-analysis, and the results of the literature search. The next chapter contains a literature review of all studies found that researched the relationship between a predictor and some measure of pilot success.

III. A Review of Applicable Literature

Introduction to the Chapter

This chapter contains a review of the literature applicable to the research presented in this thesis. It begins with the history of pilot selection, develops some factors measured to determine suitability for selection, and summarizes the cognitive studies used in the meta-analysis procedure.

History of Pilot Selection

The history of pilot selection dates back to the early 20th century. Researchers were first motivated to effectively select pilots for World War I (Davis, 1989:9). Over the years researchers have attempted to correlate psychological (personality), physiological, cognitive, psychomotor, and various biographical factors with the success of pilots in undergraduate pilot training and in operational flying.

World War I. Before World War I (WWI), the United States Army had no program for selecting pilots. In order to develop a legitimate program for selecting pilots, a group of psychologists gathered a series of psychological tests to measure pilot ability (Davis, 1989:9). Following WWI, psychologists expanded their research into these "human factors" to include the relationship to aircraft accidents. When researchers concluded that many of the accidents could be related to human error, they were encouraged to do more research. This eventually led to studying pilot selection methods to determine if it were possible to identify and measure traits in individuals that contribute to aircraft accidents. Pilot selection studies were then expanded to include traits that are associated with successful pilots.

World War II. World War II (WWII) brought about the need for a large number of qualified pilots and aircrews. The task of selecting pilots was given to the psychologists at the Army Air Force School of Aviation Medicine. Initial applicants first took the Army Air Force Qualifying Examination (AAFQE) which tested aptitude, motivation, and attitude. Those who did well on the AAFQE then took the Aircrew Classification Battery (ACB) which included 14 different tests (Cooper, 1976:6-7). The AAFQE was apparently successful in identifying traits required to succeed in the Army's pilot training program. After using the test to screen out candidates prior to entry into training, the attrition rate of individuals in the program was cut in half. (North and Griffin, 1977:10).

Rossander noted that toward the end of the war an effort was made to replace the AAFQE and ACB with less expensive, less time-consuming commercial tests. During this period, more than 20 studies were conducted with no significant relationships developed. The reason for this could stem from the fact that the tests were designed to screen out unqualified individuals, rather than predict their success (Rossander, 1980).

Pilot Selection Predictor Category Chosen for Meta-analysis

For the purpose of this research, the projects reviewed were categorized by the type of predictors used to evaluate the relationship with the successful completion of UPT. The predictors were divided into four categories; demographic, psychomotor, personality, and cognitive. Some studies, as mentioned earlier, contain information that falls into more than one category. Those studies that contained cognitive predictors are reviewed because the meta-analysis was completed on studies in this category.

Cognitive Abilities and the Prediction of Pilot Success.

Cognitive predictors refer to those characteristics that measure one's ability to process, store, perceive, encode, transform, and compare information. It is important, given the sophistication of current aircraft, for a pilot to have the ability to perform these functions with both speed and accuracy. The reader should be aware that the correlations reported here might, at first glance, look insignificant. The cognitive predictors used for the meta-analysis represent only a fraction of the predictors used for pilot selection. The objective of this study is to clarify the size and significance of cognitive predictors. Once the predictive validity of these predictors is clarified, it will allow researchers to gage the need for other predictors.

USAF Cognitive Studies. The most common cognitive test battery used by the Air Force is the Air Force Officer Qualifying Test (AFOQT). The first form of the AFOQT was developed in 1951 (Form A) by taking subtests from the Aircrew Classification Battery (ACB). The questions and form of the test are changed periodically to prevent "test compromise opportunity, and to improve the predictive validity of the battery" (Skinner et al., 1987:1).

The AFOQT consists of 16 subtests. These subtests are then combined to quantify the subject's ability on five different composites (See Table 1): verbal, quantitative, academic aptitude, pilot, and navigator-technical (Rogers et al., 1986:2).

Of the studies reviewed for this research, eleven specifically addressed the predictive validity of the AFOQT in assessing a candidate's potential for successfully completing undergraduate pilot training. Carretta conducted four of those studies, which looked at other cognitive predictors in conjunction with the AFOQT.

McGrevy and Valentine investigated the AFOQT in conjunction with two psychomotor tests. They were not able to demonstrate a significant correlation between any of the five AFOQT composites investigated (pilot, navigator-technical, officer quality, verbal, and quantitative) and success in pilot training (McGrevy and Valentine, 1974:17).

TABLE 1

AFOQT Subtests and Composites

Subtests	Pilot	Navigator Technical	Academic Aptitude	Verbal	Quantitative
	Composites				
Verbal Analogies	X		X	X	
Arithmetic Reasoning		X	X		X
Reading Comprehension			X	X	
Data Interpretation		X	X		X
Word Knowledge			X	X	
Math Knowledge		X	X		X
Mechanical Comprehension	X	X			
Electrical Maze	X	X			
Scale Reading	X	X			
Instrument Comprehension	X				
Block Counting	X	X			
Table Reading	X	X			
Aviation Information	X				
Rotated Blocks		X			
General Science		X			
Hidden Figures		X			

In a separate study, however, Bordelon and Kantor did find a relationship between the AFOQT composite, in conjunction with two psychomotor test scores, and success in pilot training. After analyzing the scores of 4,460 candidates, Bordelon and Kantor determined that the implementation of the psychomotor screening would add to the predictive ability of the five AFOQT composites included in the study. The correlation between the AFOQT Pilot Composite and the successful completion of pilot training was reported as .158. This indicates those who scored higher on the AFOQT Pilot Composite were more likely to complete UPT (Bordelon and Kantor, 1986).

In 1987, Carretta conducted separate studies on three different cognitive tests: mental rotation test (measures spatial ability), the embedded figures test (measures field dependence-independence), and a compensatory tracking and signal tracking dual-task (measures cognitive time-sharing ability). The predictive utilities of spatial ability and field dependence-independence, with respect to success in UPT, were evaluated when used alone and in conjunction with the AFOQT pilot composite.

With regard to the Mental Rotation Test (spatial ability), Carretta concluded spatial ability alone was not "useful for predicting successful completion of UPT, but was significantly related to a post-UPT advanced training recommendation" (Carretta, 1987b:7). The AFOQT pilot composite was found to have a correlation of .12 ($p<.05$), with success in UPT, and the combination of the two (spatial ability and the AFOQT pilot composite) had a correlation of .136 ($p<.05$) in a regression. Carretta concluded that those who scored higher on the AFOQT were more likely to complete UPT, and that the mental rotation test slightly improved the ability to predict completion of UPT.

The field dependence-independence ability (the ability to distinguish embedded figures) measure showed similar results. For the field dependence-independence measure, the AFOQT pilot composite had a correlation of .109 ($p<.01$), and the combination had a correlation of .126 ($p<.05$) to success in UPT. Again, the strength of the relationship between the test and completion of UPT was improved by the inclusion of the field dependence-independence measurement (Carretta, 1987a). The last cognitive test, Time-Sharing ability, failed to provide any additional predictive validity with respect to completion of UPT (Carretta, 1987c).

In 1988, Carretta administered two tests to 2,219 United States Air Force pilot candidates prior to their entry into UPT. The two tests, Encoding Speed (encoding and classification ability) and Immediate/Delayed Memory (short-term memory retrieval), were evaluated for their relationship to flight training performance. AFOQT scores were also available for these subjects. The AFOQT portion of the study found a correlation of .09 ($p < .05$) with success in UPT, indicating those who scored higher on the AFOQT were more likely to complete UPT. The contributions of the two new tests were mixed. Although both tests were found to be reliable instruments, only the results of the Encoding Speed test was significantly related to higher performance in flight training and the recommendation for additional training in a fighter, attack, or reconnaissance aircraft. Carretta also found that the combination of the AFOQT pilot composite and Encoding Speed increased the correlation from .09 ($p<.05$) to .156 ($p<.05$). This indicates that those who score higher were more likely to complete UPT, and that the Encoding Speed test improved the predictive validity with respect to completion. Immediate/Delayed Memory failed to demonstrate a significant relationship to UPT pass/fail, and did not contribute to the predictive validity of the AFOQT pilot composite (Carretta, 1988).

In 1988, Carretta and Siem conducted a study that included the results of correlations of each of the AFOQT subtests with respect to UPT outcome. A multiple R for the entire AFOQT was calculated to be .285 ($p<.0001$). These results indicate that the cognitive characteristics measured through the administration of the AFOQT provide significant predictive validity with respect to predicting successful completion of UPT (Carretta and Siem, 1988).

In 1989, Colonel Roy Davis conducted a study of using personality measures to predict the success of undergraduate pilot trainees. His measures of interest included the academic aptitude, pilot, navigator-technical, verbal, and quantitative composites of the AFOQT and the respective relationships to UPT completion. He concluded that the pilot (r=.165), navigator-technical (r=.159), and quantitative composites (r=.150) showed a significant relationship ($p<.05$) to UPT completion, indicating that candidates who scored higher on these composites were more likely to complete UPT (Davis, 1989).

Arth et al., researched the Air Force Officer Qualifying Test (AFOQT) and its applicability to predicting undergraduate pilot training success. Seven of the sixteen AFOQT subscales (identified in Table 2) were significantly ($p<.05$) related to success in UPT. The pilot, navigator-technical, and verbal composites were also significantly ($p<.05$) related to success in pilot training. The researchers concluded that the findings "support the use of specialized aircrew composites to select pilots" (Arth et al., 1990:9).

TABLE 2

AFOQT Subtests Significantly Related to UPT Success

Data Interpretation	Word Knowledge
Mechanical Comprehension	Scale Reading
Instrument Comprehension	Aviation Information
Rotated Blocks	

(Arth et al., 1990:9)

In addition, Carretta used a subject group of 885 USAF undergraduate pilot candidates randomly assigned to two groups. The two groups were used to cross-validate pilot selection models that use a combination of the Air Force Officer's Qualifying Test (AFOQT) and the Basic Attributes Test (BAT). The BAT consists of two psychomotor tests (two-hand coordination and complex coordination), four cognitive tests (encoding speed, mental rotation, item recognition, time-sharing), and two personality tests (self-crediting word knowledge and activities interest inventory). He concluded the selection models were significantly related to UPT final outcome for both groups, and that students with good hand-eye coordination and who made quick decisions (time-sharing test) were more likely to complete UPT (Carretta, 1990).

Cowan, Barrett, and Wegner examined the Air Force Officer Qualifying Test (AFOQT) and its relationship to several criterion measures (e.g., performance in Officer Training School, UPT, etc.). Using 59 possible predictors, the researchers built a model to predict each of the various criterion measures. With regard to the prediction of UPT performance, they concluded that the following factors were significantly ($p<.05$) related to UPT pass/fail: the combination of a private pilot license and the completion of calculus, the AFOQT navigator-technical composite, a military applicant (negative correlation), a civilian applicant, a commercial pilot's license, work experience (full-time, non-managerial, non-supervisory), and the recruiter's evaluation on the applicant's communication skills (Cowan et al., 1990).

United States Navy Cognitive Studies. The Navy also uses cognitive testing to screen its candidates. The Navy uses the Academic Qualification Test/Flight Aptitude Rating (AQT/FAR). The AQT/FAR is the Navy/Marine Corps aviation selection battery. The test battery is composed of four separate multiple choice tests: Academic Qualification Test (AQT), Mechanical Comprehension Test (MCT), Spatial Apperception Test (SAT), and the Biographical Inventory (BI). The AQT is a "single test that measures such attributes as general intelligence, verbal and quantitative abilities, clerical skills, and situational judgment" (Dolgin et al., 1987:482). The FAR composite is a combination of the scores on the MCT, SAT, and BI. The MCT examines the individual's ability to perceive physical relationships and solve practical problems (mechanical ability). The SAT is concerned with the candidate's ability to perceive spatial orientations. The BI evaluates characteristics such as maturity, risk-taking behavior, and level of aviation knowledge (Morrison, 1988:4).

In 1966, Peterson et al., studied the relationship between a measure of student pilot carefulness (as rated by their peers) and success in naval flight training. They also included the correlations of the AQT, MCT, SAT, and BI

(N=529). Of the four measures, all but BI (p>.05) were found to be significantly related (p<.01) to the successful completion of undergraduate flight training. These results indicate that those who scored higher on the AQT, MCT, and SAT were more likely to complete Naval Flight Training. The BI failed to provide predictive validity with respect to completion of flight training (Peterson _et al_., 1966:4).

Fleischman _et al_., also studied the relationship of the AQT, MCT, SAT, and BI to the successful completion of flight training. They, however, concluded that the BI and SAT demonstrated a significant (p<.05) relationship to the pass/fail criterion, and failed to demonstrate a significant relationship for the AQT or MCT with respect to the successful completion of flight training (N=575). These results indicate that candidates who scored higher on the BI and SAT were more likely to complete flight training, and that the score on the AQT or MCT was not related to successful completion. (Fleischman _et al_., 1966).

In 1973, W. L. Waag _et al_., studied the relationship of the MCT, AQT, SAT, and BI to the successful completion of Naval Flight Training. They failed to demonstrate a significant relationship between any of the four subtests (MCT, AQT, SAT, and BI) and the successful completion of flight training, indicating that candidate scores were not related to performance in flight training (Waag _et al_., 1973:5).

Hopson _et al._, examined the development and evaluation of a Naval Flight Officer Scoring Key for the Naval Aviation Biographical Inventory in 1978. In the process of evaluating the new key they also examined the AQT, MCT, SAT, and the BI (N=1039). All were found to be significantly related the pass/fail criterion ($p<.05$). The new BI investigated was found to be superior to the old, with the correlation changing from .156 ($p<.05$) to .376 ($p<.05$). These results indicate that a higher score on the new BI was better at selecting those candidates more likely to complete flight training than the old BI. They also indicate that candidates who scored higher on the AQT, MCT and SAT were more likely to complete flight training (Hopson et al, 1978:7).

Similarly, Griffin and Hopson studied the Omnibus Personality Inventory, AQT, MCT, SAT, and the BI and their relationship to the outcome of pilot training. They evaluated four separate groups containing a combined total of 1,108 subjects, and found all four of the selection battery subtests to be significantly related to the final outcome of pilot training (Griffin and Hopson, 1979:6-9).

In 1982, Griffin and Mosko conducted research for the Navy, evaluating two "dichotic" listening tasks for their usefulness in predicting performance in naval flight training. In addition, they evaluated the U.S Naval and Marine Aviation Selection Battery (AQT, MCT, SAT, and BI). They concluded that the selection battery was not correlated with the dichotic listening tasks. In addition, they concluded that the BI and FAR composite (combination of MCT, SAT, BI) demonstrated a significant relationship to the pass/fail criterion. The study had a sample size of 48 (Griffin and Mosko, 1982:9).

Thomas and Clipper studied the relationship between performance on a perceptual-motor task and a pen-and-paper achievement motivational test. Also included, was information on the AQT, SAT, FAR, MCT, and BI. They had a sample size (N=16), and failed to demonstrate a relationship between successful completion of flight training and any of the four subtests (Thomas and Clipper, 1983:20). This study, of course, is still useful for the meta-analysis, since small sample artifacts are corrected along with the correction of large sample artifacts.

In 1986, Griffin and McBride investigated predicting the success of undergraduate pilot candidates using a multi-task performance measure. The AQT and FAR scores that were included in the study demonstrated a significant relationship to the final flight grade received by the subject, but only the FAR composite was significantly related to the pass/fail criterion, indicating that those candidates who scored higher on the FAR were more likely to complete flight training (Griffin and McBride, 1986:8).

Dolgin et al., attempted to validate a test to measure the risk-taking tendencies of undergraduate flying students, and its ability to predict the successful completion of these Navy pilot trainees. They also looked at the AQT, MCT, SAT, BI, and FAR with respect to success in pilot training. With a sample size of 15, they failed to demonstrate any significant relationships between the four subtests and completion of naval aviation training, indicating that the score on these subtests was not predictive of the successful completion of flight training. (Dolgin et al., 1987:483).

Similar to that of Griffin and Mosko, Griffin and Collyer completed a follow-up study on the development and evaluation of an automated series of single and multiple dichotic listening (DLT) and psychomotor tasks (PMT). The cognitive test results included in the study were the AQT and the FAR. Their results indicated that those who took the forward series of the PMT and DLT did not demonstrate a relationship between the AQT/FAR and pass/fail, while those who took the backward series demonstrated a relationship ($p<.05$) between the FAR and pass/fail (Griffin and Collyer, 1987:10).

In 1988, Morrison studied complex visual information processing, the AQT, FAR, MCT, SAT, and BI aptitude scores, and their relationship to primary flight training success. With a sample size of 451 subjects, they were not able to demonstrate any relationship between the FAR and success in flight training. They were, however, able to demonstrate a relationship between the complex visual task performance ($r=-.274$) and the pass/fail criterion (Morrison, 1988:9).

Delaney sought to validate the Dichotic Listening and Psychomotor Task performance as a predictor of success in primary flight training. The research included the investigation of the FAR and AQT. He was not able to demonstrate a significant relationship between the two cognitive measurements and pass/fail criterion. Using a "statistically optimal" combination of the DLT, PMT, selection battery test scores, and various demographic variables, he felt he could identify the individuals who were "relatively" less likely to complete pilot training. The correlation between the FAR and success in pilot training was .14 with $p<.05$ (Delaney, 1990:7).

Although the studies on cognitive predictors indicate a positive ability to predict the success of undergraduate pilot trainees, there is a wide range of conclusions on exactly what those correlations are. Combining the studies and calculating one statistic for the pilot composite portion of these predictors should give a good indication of the true correlation between the composite (with associated subtests) and the success or failure of undergraduate pilot trainees (results of this meta-analysis are discussed in Chapter 4).

Chapter Summary

This chapter discusses the literature applicable to the research presented in this thesis. It begins with the history of pilot selection, identifies the categories of predictors reviewed, and summarizes those cognitive studies used in the meta-analysis process.

IV. Results

Introduction to the Chapter

This chapter presents the results obtained following the cumulation procedures outlined in chapter 2 (pg 13). It contains the data and calculations used to derive the correlation between the Air Force Officer Qualifying Test (AFOQT) Pilot Composite and the successful completion of undergraduate pilot training (UPT), and the correlation between the Flight Aptitude Rating (FAR) portion of the Navy and Marine Aviation Selection Battery (NMASB) and the successful graduation from Navy Flight Training.

The chapter is organized according to the order in which the meta-analysis corrections of the correlations were completed. It contains corrections for sampling error, error of measurement, dichotomous criterion, and restriction of range.

Tables 3 and 4 contain the data gathered from the USAF and US Navy studies used for the meta-analysis. They include; the author's name, year of the study, correlations between the predictor (AFOQT Pilot Composite or Navy FAR) and the criterion (successful completion of flight training), and the sample size for each study. The data in these tables was used for the corrections completed using the meta-analysis process.

TABLE 3

Author, Year of Study, Correlation Statistics and Sample Size of Studies Using Cognitive Predictors of the Air Force Officer Qualifying Test Pilot Composite to Predict Successful Completion of Undergraduate Pilot Training

Author	Year	r	N
Arth et al	1990	.210	695
Bordelon	1986	.158	4460
Carretta	1988	.090	545
Carretta	1987b	.120	526
Carretta	1987a	.109	602
Carretta/Siem	1988	.090	431
Davis	1988	.165	664
Hunter et al	1978	.150	245
Lemaster	1974	.160	71

TABLE 4

Author, Year of Study, Correlation Statistics and Sample Size of Studies Using Cognitive Predictors of the Navy/Marine Flight Aptitude Rating to Predict Successful Completion of Naval Flight Training

Author	Year	r	N
Delaney	1990	.140	530
Dolgin et al	1987	-.17	15
Griffin/Collyer	1987	.288	98
Griffin/Collyer	1987	-.08	105
Griffin/Mcbride	1986	.452	50
Griffin/Mosko	1982	.375	48
Morrison	1988	.149	451
Thomas/Clipper	1983	.430	16

Sampling Error

According to Hunter, Schmidt, and Jackson: "if the population correlation is assumed to be constant over studies, then the best estimate of that correlation is not the simple mean across studies, but a weighted average." (Hunter et al, 1982:40) The weighted average is calculated using the formula:

$$\bar{r}=\frac{\sum [N r_i]}{\sum N_i} \tag{4-1}$$

where r_i is the correlation in study i and N_i is the number of subjects in study i. The weighted average correlation of the AFOQT pilot composite was .149, and the Navy's FAR was .157. Accordingly, the "frequency weighted average squared error" (variance) is given by the formula;

$$S_r^2=\frac{\sum [N_i (r_i-\bar{r})^2]}{\sum N_i} \tag{4-2}$$

The frequency weighted average squared error for the AFOQT pilot composite was .000961, and the Navy's FAR was .01315. The variance measured above is a "confounding" of two things; "variation in population correlations (if there is any) and variation in sample correlations produced by sampling error" (Hunter et al., 1982:42). Hunter et al., present the following formula to estimate the population variance, corrected for sampling error:

$$\sigma_p^2=\sigma_r^2-\sigma_e^2=S_r^2-\frac{(1-\bar{r})^2 K}{N} \tag{4-3}$$

where K is the number of studies, and N is the total sample size of K studies. When applied to the research, the estimate of the population variance derived for the AFOQT pilot composite is .00083, and .00736 for the Navy's FAR.

Error of Measurement

The correction for error of measurement is given by the formula:

$$r_c = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}} \qquad (4\text{-}4)$$

where r_{xy} is the correlation between the predictor and criterion, r_{xx} is the reliability of the predictor measurement, and r_{yy} is the reliability of the criterion measurement. r_{xx} equals .964 for the AFOQT portion of this research (Rogers et al., 1986:6). For this calculation, r_{yy} is considered to be 1, since there is no reliability measure reported for the pass/fail criterion measure. This results in a conservative estimate for the corrected correlation. The corrected correlation for the Air Force sample is .152.

The reliability of the FAR was not available in any of the literature, nor was it released by the Naval Aerospace Medical Institute (the controlling organization for FAR testing). Other pilot selection researchers were contacted (D. R. Hunter, F. M. Siem, and T. R. Morrison), but did not recall the reliability being reported. Since this value could not be obtained, this correction was not done for the FAR.

Restriction of Range

Recall from earlier chapters that the population for this study (undergraduate pilot trainees) was pre-selected from a much larger population (all those who took either the AFOQT or the Naval/Marine Aviation Selection Battery). In order to correct the correlation (r_c) calculated above, it is necessary to find the ratio of the standard deviation in the population to that of the study group. This ratio is called U, given by the formula:

$$U=\frac{S}{s} \qquad (4\text{-}5)$$

where S is the standard deviation of scores for the unrestricted group (all those who took the test), and s is the standard deviation of the scores for the restricted population (those selected to attend flight training). The Air Force values for S (27.84) and s (18.76) were taken from a study conducted by Arth et al., (Arth et al., 1990:4,11). They presented the scores for a group of 3000 subjects who took the AFOQT. The calculated ratio (U) for the Air Force sample was 1.56.

Once again, the standard deviations of the restricted and unrestricted populations were neither reported nor released from the Naval Aerospace Medical Institute. Therefore, the ratio could not be calculated. This correction was not done for the FAR composite.

The correction for restriction of range is given by the formula:

$$r_c = r_o * U \quad (4\text{-}7)$$

where r_c is the uncorrected correlation. The corrected correlation is .237 for the Air Force sample, and was not calculated for the Navy sample.

Dichotomization

The correction for dichotomization is given by the formula:

$$r = \frac{r_o}{.80} \quad (4\text{-}6)$$

where r_0 is the observed correlation, and .80 is the correction due to a 50/50 split in the criterion alternatives (pass or fail). Hunter and Schmidt identify the correction for a 50/50 dichotomous variable to be .80. The corrected correlation is .296 for the Air Force sample , and .196 for the Navy sample (Hunter and Schmidt, 1990:47).

Confidence Intervals

Confidence intervals ($p < .05$) were calculated around the corrected and uncorrected correlations for both the Air Force Officer Qualifying Test and the Navy's Flight Aptitude Rating. The results are presented in Table 4-3. Calculating the confidence intervals allows for direct comparison between the Air Force and Navy studies. It also allows for the direct comparison of the ccrrected and uncorrected correlations within each of the two groups of studies (Air Force and Navy). Further discussion of these results is contained in Chapter V.

Chapter Summary

This chapter presented the results obtained following the cumulation procedures outlined by Hunter, Schmidt and Jackson. It contains the data and calculations used to derive the corrected correlation between the AFOQT Pilot Composite and success in UPT, and the correlation between the FAR and success in Naval Flight Training.

TABLE 5

Artifacts and Corrected Correlations for the Air Force and Navy Studies, and Confidence Intervals for the Corrected/Uncorrected Correlations Used to Predict Successful Completion of Flight Training

Artifact	Air Force	Navy
Sampling Error		
Weighted Avg	.149	.157
Sample Variance	.000961	.01316
Population Variance	.000083	.00736
Error of Measurement	.152	****
Restriction of Range	.237	****
Dichotomization	.297	.196

Confidence Intervals ($p<.05$)		
Uncorrected	$.088< r <.210$	$-.068< r <.382$
Corrected	$.279< r <.315$	$.027< r <.365$(*)

(*) Corrected correlation does not include correction for error of measurement and restriction of range (Necessary data was not available).

V. Discussion and Conclusions

Introduction to the Chapter

This chapter covers the findings of the meta-analysis calculations presented in Chapter IV. It concludes with an examination of the hypothesis presented in chapter I of this study.

Discussion

Air Force Officer Qualifying Test Pilot Composite.

The meta-analysis procedure indicates that there is a measurable correlation between the pilot composite of the AFOQT and completion of pilot training. The uncorrected weighted mean correlation between the pilot composite scores and completion of UPT was $r=.149$ (nine studies were included in the meta-analysis calculations). A 95 percent confidence interval was calculated to be $.088 < r < .210$ for the uncorrected correlation. Indicating that the correlation is statistically significant ($p<.05$) since the interval does not include zero.

When the correlation was corrected for sampling error, error of measurement, restriction of range, and dichotomization, the correlation was increased to .297, with a 95 percent confidence interval of $.279 < r < .315$. Comparison of the corrected and uncorrected intervals indicated that the meta-analysis procedure made a significant ($p<.0001$) difference in improving the magnitude of the correlation.

Navy/Marine Aviation Selection Battery (NMASB).

The same meta-analysis procedures was conducted on eight Naval studies involving the use of the Flight Aptitude Rating (FAR) to predict the success of undergraduate pilot trainees. Since the FAR is designed to measure the same abilities as the pilot composite of the AFOQT, it was expected that the results of the meta-analysis would be similar. The weighted mean correlation was found to be $r=.157$. Although this was close to the correlation for the AFOQT pilot composite, there were some statistically significant differences between the two groups of studies (between the Air Force and Navy). Because the Navy studies included a higher degree of variance, calculation of the 95 percent confidence interval included zero in its range ($-.068< r <.382$). Therefore, it could not be concluded that the uncorrected correlation of the FAR was significant in predicting completion of Naval Flight Training.

The corrected correlation (corrected for sampling error and dichotomization) was .196, with a 95 percent confidence interval of $.027< r <.365$. Comparison of the corrected and uncorrected correlations indicated that the meta-analysis did not have a statistically significant effect on improving the magnitude of the correlation. This is due to the fact the corrections for error of measurement and restriction of range could not be performed, since the Naval Aerospace Medical Institute would not release the data required to perform these calculations. However, it is unlikely that the corrected correlation would have been significantly different from the uncorrected, due to the high variance of results present in the Navy studies. In order for the difference between the corrected and uncorrected correlations to be statistically significant ($p<.05$), the corrected correlation would have to be corrected to a magnitude of approximately .55 (equals the upper confidence limit of the uncorrected plus a z-score of 1.96 times the standard deviation of the population). The research indicates that .55 is probably beyond the predictive validity of a standardized cognitive test for the prediction of a dichotomous criterion.

Conclusions

The present study has demonstrated important evidence for using both the Air Force Officer Qualifying Test Pilot Composite and Navy/Marine Flight Aptitude Rating as selection devices for their respective pilot training. The following are the main conclusions of this study. They are followed by the research hypotheses of Chapter I.

General Conclusions. The following are the eight general conclusions reached as a result of this study:

1. There is an identifiable and statistically significant ($p<.0001$) uncorrected correlation between the

Air Force Officer Qualifying Test Pilot Composite and the successful completion of undergraduate pilot training.

2. There is an identifiable and statistically significant ($p<.0001$) corrected correlation between the Air Force Officer Qualifying Test Pilot Composite and the successful completion of undergraduate pilot training.

3. There is an identifiable, but not statistically significant ($p<.05$) uncorrected correlation between the Navy/Marine Aviation Selection Battery Flight Aptitude Rating and the successful completion of Naval Flight Training.

4. There is an identifiable and statistically significant ($p<.05$) corrected correlation between the Navy's Flight Aptitude Rating and the successful completion of Naval Flight Training.

5. There is no statistically significant ($p<.05$) difference, with respect to the successful completion of flight training, between the uncorrected correlation of the Air Force Officer Qualifying Test Pilot Composite and the uncorrected correlation of the Navy Flight Aptitude Rating.

6. There is no statistically significant difference, with respect to the successful completion of flight training, between the corrected correlation of the Air Force Officer Qualifying Test Pilot Composite and the partially corrected correlation of the Navy's Flight Aptitude Rating.

7. There is a statistically significant difference between the corrected and uncorrected correlations for the Air Force Officer Qualifying Test Pilot Composite.

8. There is no statistically significant difference between the uncorrected and partially corrected correlations of the Navy Flight Aptitude Rating. As stated earlier, it is unlikely that the additional data would have made this difference statistically significant. The magnitude of the corrected correlation would have to be approximately .55, and the literature indicates that this magnitude is probably beyond the predictive validity of a standardized cognitive test used for the prediction of a dichotomous criterion.

Test Hypotheses. The test hypothesis for the Air Force sample is as follows:

Ho: The mean corrected correlation between the AFOQT Pilot Composite and success in undergraduate pilot training is not statistically significant ($p<.05$).

Ha: The mean corrected correlation between the AFOQT Pilot Composite and success in undergraduate level is statistically significant ($p<.05$).

The test hypothesis for the Navy sample is as follows:

Ho: The mean corrected correlation between the Navy Flight Aptitude Rating and success in flight training is not statistically significant ($p<.05$).

Ha: The mean corrected correlation between the Navy Flight Aptitude Rating and success in flight training is statistically significant ($p < .05$).

Research Hypotheses. The research hypothesis for both the United States Air Force and United States Navy samples is the alternative hypotheses stated above. The research indicates that the mean corrected correlations for both the Air Force Officer Qualifying Test Pilot Composite and Navy Flight Aptitude Rating is statistically significant. Both null hypotheses are rejected because the calculated 95 percent confidence interval did not include zero. Therefore, the alternative hypotheses are accepted.

Chapter Summary

This chapter contains the findings of Chapter IV and conclusions resulting from the meta-analysis procedure. The following chapter will address the recommendations of the researcher.

VI. Recommendations

Introduction to the Chapter

This chapter contains recommendations for future research in this area. These recommendations are made in light of the present study.

Recommendations

The following are recommendations based on the research conducted for this study:

1. First, it is recommended the Department of Defense (DOD) conduct a study on the reporting procedures of DOD related research. Many difficulties encountered in this research could be directly attributed to a lack of rigor and lack of completeness in DOD sponsored research. The cumulative effects (realized by follow-on research) of these shortcomings are unknown, but potentially significant. A study on the procedures used in DOD sponsored research could at least identify specific shortcomings and make recommendations for improvement. The study could first look at whether or not DOD related studies generally report the same statistics. Many of the studies in this research only contained the correlations derived through computer programs, and did not report standard deviations or variances. The study could also include a survey of current DOD researchers on information they feel is important to include.

Adoption of this recommendation would take DOD one step closer to a research standard that would make reviewing by future researchers much easier and meaningful. This type of standard would also make meta-analysis of certain topic areas much easier.

2. It is recommended that future research look at other aspects of pilot training and perform a similar meta-analysis. One or more of the other categories of characteristics (psychomotor tasks, personality, or demographics) should be studied in order to broaden the analysis of pilot selection research. In addition, studies should continue on more specific subelements of the categories (such as two-hand coordination as a subset of psychomotor tasks).

3. It is recommended that future research address the selection methods used by other countries. This could include both the pilot selection tests and flight screening portions of the programs. As the analytical power of meta-analysis lies in the error-canceling effects of comparing multiple studies, an effort should be made to compare and contrast all aspects of each program and the associated advantages and disadvantages of each. The attrition rates of each program, along with any moderating variables should also be developed. Research along these lines will further isolate the most important traits contributing to success in pilot training.

4. It is recommended that future research include administering Navy/Marine Aviation Selection Battery to Air Force pilot candidates. This would be in addition to the Air Force tests already administered, and would provide an experimental control to directly compare the two tests. For example, the researcher could compare the predictive validity of both selection batteries on the final grade of the flight screening program, or the T-37 phase of undergraduate pilot training. The Air Force Human Resources Laboratory would be a good sponsor for this type of research.

5. It is recommended that future research include administering a demographic, psychomotor, personality, or cognitive battery to a group of established pilots prior to scored bombing runs. This could be done before a bombing competition, or for normal scored wing level bombing runs. The results of the test could then be compared to the successful delivery of weapons, given a certain type of aircraft. None of the criteria studied previously mean more than the ability to put a bomb on target. A study of this sort might shed some light on what distinguishes the good combat pilot.

Chapter Summary

This chapter concluded the research effort. It contains recommendations made for future research in light of conclusions made during this research project. The findings of this research indicate that both the Air Force Officer Qualifying Test Pilot Composite and Navy/Marine Flight Aptitude Rating are useful in selecting those candidates who are more likely to complete pilot training.

Bibliography

Arth, T. O., K. W. Steuk, C. T. Sorrentino, and E. F. Burke. Air Force Officer Qualifying Test (AFOQT): Predictors of Undergraduate Pilot Training and Undergraduate Navigator Success (AFHRL-TP-89-52, AD-A221 674). Brooks AFB TX: Air Force Human Resources Laboratory, 1990.

Bordelon, V. P., and J. E. Kantor. Utilization of Psychomotor Screening for USAF Pilot Candidates: Independent and Integrated Methodologies (AFHRL-TR-86-4, AD-A170 353). Brooks AFB TX: Air Force Human Resources Laboratory, 1986.

Carretta, T. R. Cross-Validation of Experimental USAF Pilot Training Performance Models (AFHRL-TR-89-68, AD-A222 253). Brooks AFB TX: Air Force Human Resources Laboratory, 1990.

Carretta, T. R. Field Dependence-Independence and its Relationship to Flight Training Performance (AFHRL-TP-87-36, AD-A188 888). Brooks AFB TX: Air Force Human Resources Laboratory, 1987a.

Carretta, T. R. Relationship of Encoding Speed and Memory Tests to Flight Training Performance (AFHRL-TP-87-49, AD-A190 670). Brooks AFB TX: Air Force Human Resources Laboratory, 1988.

Carretta, T. R. Spatial Ability as a Predictor of Flight Training Performance (AFHRL-TP-86-70, AD-A183 141). Brooks AFB TX: Air Force Human Resources Laboratory, 1987b.

Carretta, T. R. Time-Sharing Ability as a Predictor of Flight Training Performance (AFHRL-TP-86-69, AD-A181 838). Brooks Air Force Base, Texas.: Air Force Human Resources Laboratory, 1987c.

Carretta, T. R., and F. K. Siem Personality, Attitudes, and Pilot Training Performance: Final Analysis (AFHRL-TP-88-23, AD-A199 893). Brooks AFB TX: Air Force Human Resources Laboratory, 1988.

Collender, S. E. The Guide to the Federal Budget. Washington DC: The Urban Institute Press, 1991.

Cooper, M. F. "Application of Operational Pilot Selection Criteria." Research study, Maxwell AFB AL: Air Command and Staff College, 1976.

Cowan, D. K., L. E. Barrett, and T. G. Wegner. Air Force Officer Training School Selection System Validation (AFHRL-TR-89-65, AD-A225 245). Brooks AFB TX: Air Force Human Resources Laboratory, 1990.

Crowley, J. S. Cerebral Laterality and Handedness in Aviation: Performance and Selection Implications (USAFSAM-TP-88-11, AD-A206 196). Brooks AFB TX: USAF School of Medicine, 1989.

Davis, Carl, and Robert Steel. Meta-Analysis: A Technique for Performing Conclusive Literature Reviews. Class Handout (COMM 630), Wright-Patterson AFB OH: Air Force Institute of Technology, 1991.

Davis, R. A. Personality: Its Use in Selecting Candidates for US Air Force Undergraduate Pilot Training (AD-A206 051). Maxwell AFB AL: Airpower Research Institute, 1988.

Delaney, H. D. Validation of Dichotic Listening and Psychomotor Task Performance as Predictors of Primary Flight Training Criteria: Highlighting Relevant Statistical Issues (NAMRL-1357, AD-A227 010). Pensacola NAS FL: Naval Aerospace Medical Research Laboratory, 1990.

Dolgin, D. L., R. N. Shull, and G. D. Gibb. Risk Assessment and the Prediction of Student Pilot Performance (AD-A201793). Pensacola NAS FL: Naval Aerospace Medical Research Laboratory, 1987.

Eisen, Major Stefan. USAF Flight Screening: First Step on the Road to Wings (AD-A192 613). Maxwell AFB AL: Air Command and Staff College, 1988.

Emory, W. C. Business Research Methods. Homewood IL: Richard C. Irwin, Inc., 1980.

Fleischman, H. L., and others. The Relationship of Five Personality Scales to Success in Naval Aviation Training (NAMI-968, AD-636 268). Pensacola NAS FL: Naval Aerospace Medicine Institute, 1966.

Fulkerson, S. C., S. L. Freud, and G. H. Raynor. "The Use of the MMPI in the Psychological Evaluation of Pilots," Journal of Aviation Medicine, 29: 122-129 (1958).

Glass, G. V., and others. Meta-Analysis in Social Research. Beverly Hills CA: Sage Publications, 1982.

Griffin, G. R., and P. D. Collyer. Development and Evaluation of an Automated Series of Single- and Multiple-Dichotic Listening and Psychomotor Tasks (NAMRL-1336, AD-A199 490). Pensacola NAS FL: Naval Aerospace Medical Research Laboratory, 1987.

Griffin, G. R., and J. A. Hopson. An Evaluation of the Omnibus Personality Inventory in the Prediction of Attrition in Naval Aviation Training (NAMRL-1253, AD-A065001). Pensacola NAS FL: Naval Aerospace Medicine Research Laboratory, 1978.

Griffin, G. R., and D. K. McBride. Multitask Performance: Predicting Success in Naval Aviation Primary Flight Training (NAMRL-1316, AD-A168 246). Pensacola NAS FL: Naval Aerospace Medical Research Laboratory, 1986.

Griffin, G. R., and J. D. Mosko. Preliminary Evaluation of Two Dichotic Listening Tasks as Predictors of Performance in Naval Aviation Training (NAMRL-1287, AD-A127 337). Pensacola NAS FL: Naval Aerospace Medicine Research Laboratory, 1982.

Henmon, V. A. C. "Air Services Tests of Aptitude for Flying," Journal of Applied Psychology, 3: No. 2 (1919).

Hopson, J. A., G. R. Griffin, N. E. Lane, and R. K. Ambler. Development and Evaluation of a Naval Flight Officer Scoring Key for the Naval Aviation Biographical Inventory (NAMRL-1256, AD-A141-523). Pensacola NAS FL: Naval Aerospace Medical Research Laboratory, 1978.

Hunter J. E., and F. L. Schmidt. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. Newbury Park CA: Sage Publications, 1990.

Hunter, J. E., F. L. Schmidt, and G. Jackson. Meta-Analysis: Cumulating Research Findings Across Studies. Beverly Hills, CA: Sage Publications, 1982.

McGrevy, D. F., and L. D. Valentine. Validation of Two Aircrew Psychomotor Tests (AFHRL-TR-74-4, AD 777 830). Brooks AFB TX: Air Force Human Resources Laboratory, 1974.

Morrison, T. R. Complex Visual Information Processing: a Test for Predicting Navy Primary Flight Training Success (NAMRL-1338, AD-A200 394). Pensacola NAS FL: Naval Aerospace Medical Research Laboratory, 1988.

North, Robert A., and Glenn R. Griffin. Aviator Selection 1919-1977 (Special Report 77-2). Pensacola NAS FL: Naval Aerospace Medical Research Laboratory, 1977.

Peterson, F. E., N. E. Lane, and R. K. Ambler. Carefulness Peer Ratings as a Predictor of Success in Naval Aviation Training (Special Report 66-1, AD 631 752). Pensacola NAS FL: Naval School of Aviation Medicine, Naval Aerospace Institute, 1966.

Rogers, D. L., B. W. Roach, and T. G. Wegner. Air Force Qualifying Test Form O: Development and Standardization (AFHRL-TR-86-24, AD-A172 037). Brooks AFB TX: Air Force Human Resources Laboratory, 1986.

Rossander, Pauline. Personality Inventories and Prediction of Success in Pilot Training: State of the Art. Willowdale, Ontario, Canada: Canadian Forces Applied Research Unit, 1980.

Siem, F. M., T. R. Carretta, and T. A. Mercante. Personality, Attitudes, and Pilot Training Performance: Preliminary Analysis (AFHRL-TP-87-62, AD-A199 983). Brooks AFB TX: Air Force Human Resource Laboratory, 1988.

Skinner, J., and M. J. Ree. Air Force Officer Qualifying Test (AFOQT): Item and Factor Analysis of Form O (AFHRL-TR-86-68, AD-A184 975). Brooks AFB TX: Air Force Human Resources Laboratory, 1987.

Thomas, G. R., and R. W. Clipper. Performance Consistency on a Perceptual-Motor Task as a Correlation of Achievement Motivation (NAMRL-1299, AD-A135 933). Pensacola NAS FL: Naval Aerospace Medical Research Laboratory, 1983.

Waag, W. L., R. H. Shannon, and R. K. Ambler. The Use of Confidential Instructor Ratings for the Prediction of Success in Naval Undergraduate Pilot Training (NAMRL-1175, AD-757 693). Pensacola NAS FL: Naval Aerospace Medicine Research Laboratory, 1973.

Vita

Captain William E. Lynch, USAF, was born 13 October, 1963, and grew up in Easton, Pennsylvania. He graduated from high school in 1981, and received a Bachelor of Science in Electrical Engineering from Wilkes University in 1986. He was commissioned on 18 May 1986 through the Air Force Reserve Officer Training Corp. From 1987 to 1990 Captain Lynch served as a Munitions Maintenance Officer at Carswell Air Force Base, Texas. He received a Master of Liberal Arts degree from Texas Christian University in 1990, and entered the School of Systems and Logistics, Air Force Institute of Technology, Air University, in May 1990.

Permanent Address: 5172 Cobb Drive
Dayton, OH 45431

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1991	Master's Thesis

4. TITLE AND SUBTITLE

A META-ANALYSIS OF PILOT SELECTION TESTS: SUCCESS AND PERFORMANCE IN PILOT TRAINING

5. FUNDING NUMBERS

6. AUTHOR(S)

William E. Lynch, Captain, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology, WPAFB OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GLM/LSM/91S-44

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)* The purpose of this study was to determine if the characteristics measured by the Air Force Officer Qualifying Test Pilot Composite and Navy/Marine Flight Aptitude Rating were significantly correlated to the successful completion of flight training. Meta-analysis was used to: calculate the mean weighted average correlation, and correct for sampling error, error of measurement, restriction of range, and dichotomization. Over 200 studies were considered for the meta-analysis. The results indicate that both the uncorrected and fully corrected weighted mean correlations for a group of nine Air Force studies were statistically significant ($p<.0001$). The partially corrected (sampling error and dichotomization) correlation for a group of eight Navy studies was also statistically significant ($p<.03$), while the uncorrected weighted mean correlation was not significant ($p>.05$). There was no significant difference between the magnitude of the correlations (corrected and uncorrected) between the Navy and Air Force groups. The findings of this research indicate that both the Air Force Officer Qualifying Test Pilot Composite and Navy/Marine Flight Aptitude Rating are useful in selecting those candidates who are more likely to complete pilot training.

14. SUBJECT TERMS

AFOQT, Navy/Marine FAR, Pilot Selection, Meta-analysis, Cognitive

15. NUMBER OF PAGES 73

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

AFIT Control Number AFIT/GLM/LSM/91S-44

AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to determine the potential for current and future applications of AFIT thesis research. Please return completed questionnaires to: AFIT/LSC, Wright-Patterson AFB OH 45433-6583.

1. Did this research contribute to a current research project?

   a. Yes b. No

2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not researched it?

   a. Yes b. No

3. The benefits of AFIT research can often be expressed by the equivalent value that your agency received by virtue of AFIT performing the research. Please estimate what this research would have cost in terms of manpower and/or dollars if it had been accomplished under contract or if it had been done in-house.

   Man Years ____________ $____________

4. Often it is not possible to attach equivalent dollar values to research, although the results of the research may, in fact, be important. Whether or not you were able to establish an equivalent value for this research (3 above), what is your estimate of its significance?

   a. Highly Significant b. Significant c. Slightly Significant d. Of No Significance

5. Comments

Name and Grade ____________ Organization ____________

Position or Title ____________ Address ____________