USE OF MULTIVARIATE TECHNIQUES TO VALIDATE AND IMPROVE THE CURRENT USAF PILOT CANDIDATE SELECTION MODEL

THESIS

Ross A. Keener, Captain, USAF

AFIT/GOR/ENS/03-13

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GOR/ENS/03-13

USE OF MULTIVARIATE TECHNIQUES TO VALIDATE AND IMPROVE THE CURRENT USAF PILOT CANDIDATE SELECTION MODEL

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Ross A. Keener, BA, M.Ed.

Captain, USAF

March 2003

AFIT/GOR/ENS/03-13

USE OF MULTIVARIATE TECHNIQUES TO VALIDATE AND IMPROVE THE CURRENT USAF PILOT CANDIDATE SELECTION MODEL

Ross A. Keener, BA, M.Ed.
Captain, USAF

Approved:

/s/_____ _____
Kenneth W. Bauer, Ph.D. (Co-Chairman) Date

/s/_____ _____
Stephen P. Chambal, Capt, USAF (Co-Chairman) Date

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Dr. Bauer, for his guidance throughout the course of this thesis effort. The insight and knowledge he contributed were instrumental. Captain Stephen Chambal also provided much needed advice and input necessary to keep this research going in the right direction. Also I would like to acknowledge the help I received from Lt Matthew Cooper and TSgt Leonard Burkhardt from Air Education Training Command Studies and Analysis Squadron for the support provided in working with the data used in this thesis.

I am also indebted to my fellow AFIT students who showed much patience in helping me along the way. I am thankful for the blessing a lovely wife who has been my encouragement throughout the entire AFIT experience. Finally, I would like to thank my Lord and Saviour, Jesus Christ, without whom none of this would be possible.

Ross A. Keener

Table of Contents

Page

Page

Page

List of Figures

Page

Page

Page

List of Tables

Page

Page

Abstract

Training pilots for the USAF costs millions of dollars every year. There are seven points of entry into Air Force Undergraduate Pilot Training (UPT). Each source has its own selection process to screen candidates accepted into UPT. The Pilot Candidate Selection Method (PCSM) seeks to ensure the highest possible probability of success at UPT. PCSM applies regression weights to a candidate's Air Force Officer Qualification Test (AFOQT) Pilot composite score, self-reported flying hours, and five Basic Attributes Test (BAT) score composites. PCSM scores range between 1 and 99 and are loosely interpreted as a candidate's probability of passing UPT.

The goal of this study is to apply multivariate data analysis techniques to validate PCSM and determine appropriate changes to the model's weights. Performance of the updated weights is compared to the current PCSM model via Receiver Operating Curves (ROC). In addition, two independent models are developed using multilayer perceptron neural networks and discriminant analysis. Both linear and logistic regression is used to investigate possible updates to PCSM's current linear regression weights. An independent test set is used to estimate the generalized performance of the regressions and independent models. Validation of the current PCSM model demonstrated in the first phase of this research is enhanced by the fact that PCSM outperforms all other models developed in the research.

USE OF MULTIVARIATE TECHNIQUES TO VALIDATE AND IMPROVE THE CURRENT USAF PILOT CANDIDATE SELECTION MODEL

I. Introduction

1.1 General Issue

The cost of the initial phase of Undergraduate Pilot Training (UPT), T-37 and T-38 phases, is approximately $137,446 per trainee (AFI 65-503, 2001). It is essential, both fiscally and operationally, that those candidates selected for UPT successfully complete training. The Air Education Training Command (AETC) currently uses the Air Force Officer Qualification Test (AFOQT) Pilot composite score, five Basic Aptitude Test (BAT) scores, and the number of self-reported Federal Aviation Administration (FAA) flying hours as inputs to the Pilot Candidate Selection Method (PCSM). PCSM is one factor considered in the selection of UPT candidates. Currently, the Air Force Academy (AFA) is the only selection source that does not incorporate PCSM. The PCSM score is a weighted linear composite of the seven predictors described above. The weights in the linear combination are regression based. The linear combination is then transformed by a discrete approximation to a sigmoidal function. It is unclear whether the current PCSM model is based on linear or logistic regression.

Recent work completed by Capt Ian Young at the Air Force Institute of Technology, Wright-Patterson AFB, OH, provided a pilot candidate selection model that apparently outperforms the current method employed by AETC (Young et al., 2003);

however, the performance is not demonstrated on a wholly independent data set. There are several drawbacks that prevent the operational implementation of Capt Young's model. These include the use of predictors not deemed feasible by AETC based on actual or perceived discriminatory affects such predictors would have on certain sub-groups of the applicant population. Despite equivalence in the model output and interpretation, implementation is further hampered because the model's format would be foreign to the selection boards.

This research focuses on three main objectives. First, given the most up-to-date data provided by AETC/SAS, the current PCSM model is validated to ensure the current predictors continue to provide optimal validity for predicting the UPT pass/fail criterion. Second, logistic and linear regressions are used to investigate possible updates to the current PCSM regression weights. Finally, an independent prediction model is developed. Retention of a similar format and interpretation as PCSM is an objective of this model; however, it is derived using other multivariate data analysis techniques for improving predictive capability. The new model may include, but is not limited to, the predictors included in the current PCSM model. This will afford AETC the ability to identify candidates with the best probability of success during pilot training, while not changing the "look and feel" selection boards are accustomed to with the current PCSM model.

1.2 Background

Approximately half of nearly 1,500 non-Air Force Academy UPT applicants are selected for pilot training each year by selection boards (Young et al., 2003). Although

not the sole basis for UPT selection, the PCSM score is used as a numeric discriminator between applicants. Each UPT candidate that attends, but does not complete the initial phase of UPT training costs the United States Air Force an estimated $72,572 (AFI 65-503, 2001). Candidates are eliminated for any of seven reasons: flying deficiency, academic deficiency, military training, medical, fear of flying, self-initiated elimination, and "other reasons".

The method for determining the PCSM score, a numeric value of 1-99, is based on a linear regression of the seven predictors; AFOQT-Pilot score, FAA flying hours, and 5 BAT scores. These predictors were selected for inclusion in PCSM based on psychometric selection theory, and many studies based on correlations sponsored by the Air Force Research Laboratory's (AFRL) Human Effectiveness Directorate and the Armstrong Laboratory's Human Resources Directorate. PCSM was commissioned in 1985 and became operational in 1993 (Ness, 1996). AETC sets minimum qualifying Air Force Officer Qualification Test (AFOQT) scores for UPT applicants by selection source. The current minimum AFOQT qualifying scores are presented in Table 1 (Carretta, 2000). PPL is an acronym for private pilot's license.

Table 1. Minimum AFOQT Qualifying Scores By Commissioning Source

Source	Pilot	Nav-Tech	P + N	Verbal	Quantitative
OTS w/ PPL	25	10	50	15	10
OTS w/o PPL	50	50	60	15	10
ROTC w/ PPL	50	10	60	15	30
ROTC w/o PPL	25	10	50	15	10
Active Duty	25	10	50	none	none
AFA (pre-1998)	none	none	none	none	none

1.3 Problem Statement

AETC provided the most up-to-date data on UPT candidate performance for use in this research. Validation of the current PCSM model and the development of an independent model that predicts UPT performance are sought using multivariate data analysis techniques. The model is validated on a wholly independent data set, the "TEST" set, to determine predictive accuracy and overall capabilities.

1.4 Research Objectives

The ultimate goal of this research is to validate the current PCSM model as well as provide AETC SAS with an improved model, which is operationally implementable within the current selection process framework. Implementation of such a model is expected to reduce UPT attrition rates and thus greatly reduce the costs associated with attrition. Once the model is developed, its performance is validated on the TEST set in comparison to PCSM.

1.5 Research Methodology

The fundamental research methodology involves the use of multivariate data analysis techniques. PCSM validation in this research is primarily accomplished via a combination of factor analysis and stepwise linear regression, partial correlations, and correlations corrected for range restriction. The magnitude of correlations (validities) is viewed as a method of quantifying a set of predictor's explanatory power for a criterion of interest. There are many statistically based complexities that tend to confound studies

based on correlations. Statistical artifacts such as range restriction, unreliability, dichotomization of criteria, group effects, factor invariance, and construct interpretation are considered. Linear and logistic regressions are used to investigate updated weights for the 7 inputs in the current PCSM model.

Three main techniques are utilized in development of an independent model. First, neural networks are investigated for their theoretically unlimited function approximating power. Second, discriminant Analysis attempts to classify an individual into a particular category based on independent predictor variables. The categories will be pass/fail for the T37 phase of UPT. Classification thresholds for the model output can be set appropriately depending on the need to maximize identification of potential failures (probability of target detection) vs. minimizing the cost/risk of false alarms. Each technique is validated for accuracy, capabilities, and limitations on the independent TEST set. Available predictors include current PCSM scores, its inputs, all AFOQT composites, BAT sub-test scores, and other demographic or quantitative variables. Although PCSM is theoretically a valid predictor, it is not considered for inclusion in the independent model. This is because the research purpose is to develop a replacement for PCSM.

1.6 Scope of Research

This research is limited to data on individuals who have been selected for UPT and have received a PCSM score since its operational implementation in 1993. Non-Air Force active duty officers are included in the data. This research is limited to three areas. First, validation of the predictors or the latent constructs underlying those predictors

currently included in PCSM. Second, updating the regression weights applied to the current PCSM model predictors. Finally, developing an independent model whose output is interpreted as a probability of passing UPT or otherwise quantifies the likelihood of success.

1.7 Outline of Thesis

This thesis is divided into the following five chapters: Introduction, Literature Review, Methodology, Findings and Analysis, and Conclusions. A brief description of each follows.

Chapter 1: Introduction – This chapter discusses the background, focus of research, research objectives, and relevance of this thesis document.

Chapter 2: Literature Review – This chapter begins with a description of PCSM and methods used to validate PCSM. Methodological issues in ability research are discussed. The pilot selection process is then reviewed. Finally, a review of multivariate data analysis techniques used in this research is presented.

Chapter 3: Methodology – The methodology chapter begins by describing the steps taken to prepare the data for analysis followed by a description of the software used in the analysis. The methodology employed in the PCSM validation study is discussed. Next, the algorithms employed in the development of an independent PCSM model are reviewed. Finally, methodologies utilized to compare the newly developed models to the current PCSM model are discussed.

Chapter 4: Findings and Analysis – This chapter presents the results of the PCSM validation study, the updated PCSM models with new weights, and the new independent

model. The validity of the models is discussed and model performance is compared to PCSM.

Chapter 5: Conclusions and Recommendations – The research results are briefly reviewed. The relevance of the research effort is presented. Recommendations for further research are provided.

II. Literature Review

2.1 Introduction to PCSM Research

The purpose of this chapter is to provide a thorough review of literature relevant to validation studies for ability research and predictive models, the Air Force PCSM model, and multivariate analysis methods. First, this chapter provides a description of the terms and issues involved with ability research validation. Second, the PCSM model is introduced. Third, methodological issues related to ability research are addressed. Fourth, this chapter provides information on the research accomplished in the development and validation of the data used in the current PCSM model as well as research relevant to pilot training selection. Additionally a discussion of the data available for analysis is presented. Finally, this chapter reviews current multivariate analysis techniques used in the analysis of the data.

2.1.1 Validity in Predictive Research

Formal validation of a predictive model is required to determine the utility of the constructs being measured and methods of recording those measurements "for predicting training and job performance" (Carretta & Ree, 2000a). "The Pearson *r* (correlation) is often used as an index of validity in psychological and educational measurement and is particularly useful when the criterion and predictor have a bivariate normal distribution" (Duan, et al., 1997). The terms validity, validity coefficient, and correlation are used interchangeably in the literature for the Pearson correlation.

A partial correlation is a correlation between a predictor and criterion conditioned on a set of predictors previously accounted for in a model. A predictor's partial correlation is dependent on the predictors included in the previously accounted for set, hence the conditional statement. When used, partial correlations are identified in this research. "The idea of partial correlation can be subsumed under 'mediation,' which means that one variable acts through another variable to exert its influence on a third variable" (Carretta & Ree, 2000a). Partial correlations provide the ability to "partial out" the influence of a set of predictors from the relationship between the criterion and another set of predictors. Ree, Carretta, & Teachout (1995) examined the influence of general cognitive ability and prior job knowledge on the acquisition of job knowledge acquired during different phases of pilot training. Lord & Novick (1968) provide the following matrix notation in Equation 1 for computing partial correlations from a matrix containing the criterion and a set of predictor scores.

$$A = \sqrt{\left(Diag \parallel \sigma^{ij} \parallel\right)} * \parallel \sigma^{ij} \parallel * \sqrt{\left(Diag \parallel \sigma^{ij} \parallel\right)} \qquad \textbf{(1)}$$

where $\|\sigma^{ij}\|$ is the inverse of the covariance matrix of the criterion and predictors and Diag(M) is a diagonal matrix of the elements on the diagonal of a matrix M. Equation 1 results in $a_{ii} = 1$ and $a_{ij} = -\rho_{ij \cdot kl\ldots}$ (partial correlation), for $k, l \ldots \neq i \neq j$. However, SPSS 11.5 is used to calculate all partial correlations in this research.

Carretta & Ree (2000a) review three historical types of validity; content, construct, and criterion (predictive). Carretta & Ree (2000a) provide the following

definitions. A predictor has content validity if it "clearly represents a knowledge area, skill, or ability." Construct validity implies the "attempt to scientifically determine what the predictor actually measures." Constructs are often given names or interpretations by the researcher. General cognitive ability is a construct often referred to, but in reality general cognitive ability is simply the name that has been attached to an abstract measure. Criterion or predictive validity is the "ability of a predictor to predict performance on an external activity or criterion."

Once suitable predictors and criteria are selected, predictive validity is examined via correlations between the predictor scores and criterion measure(s). The magnitude and direction of the correlations determine a predictor's predictive validity. In ability research for predicting success in job training, validity coefficients are usually in the small to medium range. Predictive validity can be investigated with either a predictive or concurrent validation design. In the predictive design, the criterion is only measured for those subjects selected from the sample based on performance on the predictors. This research seeks to confirm the predictive validity of the individual predictor variables currently used in the PCSM model as well as the predictive validity of the entire model.

Several predictors in a model can measure a common construct. Construct is the statistical term used to describe what is actually being measured. The term is most commonly used in factor analysis, where the underlying latent construct is interpreted through inspection of factor loadings in a factor analysis. Factor analysis is discussed in section 2.6. A very common construct within the arena of ability research is that of general cognitive ability, known simply as "*g*."

When several variables are valid predictors of a common construct, the most significant predictor is used as the baseline for calculating incremental validities of other predictors in the model. Incremental validity is the increase in predictive accuracy obtained by a predictor beyond that already accounted for with a current set of predictors (Carretta & Ree, 1994). Incremental validity is measured using partial correlations. However, in the studies reviewed for this research, incremental validity is generally conditioned on the most significant predictor alone, not a set of predictors.

Historical selection theory and studies show that measures of "g" or "psychometric *g*," are the best predictors of job/training performance. Waldman & Avolio (1989) summarize Gottfredson's (1986) and Hunter & Hunter's (1984) conclusions that "general cognitive ability not only predicts job performance moderately well but does so better than tests of any other single attribute." Morales & Ree (1992) found that general cognitive ability "was a better predictor than specific abilities or job knowledge" for prediction of 5 pilot related criteria (Carretta & Ree, 1994).

In a study on the influence of job complexity on the validity of general cognitive ability, Jones & Ree (1998) show that "job ability differences did not moderate (affect) the relationship between the amount of *g* measured by a test and its score validity." This directly refutes the *specificity doctrine*, which hypothesizes that valid predictors of one job will not be valid for others (Jones & Ree, 1998). Hence, Jones & Ree contend that for a test that measures a certain amount of *g*, that test can be expected to perform equally well at predicting job training success across a wide range of job skills or job complexity.

One concern related to validation studies is the stability of the validity coefficient across a test's score range. In a study of 68,672 Navy recruits, Lee & Foley (1986)

argued that the corrected (for range restriction) "validity coefficient is not a constant value for all subjects, rather a varied degree dependent on the level of the predictor score." They suggest treating "validity, (regression) slope, and standard error of estimate as an average rather than a constant value for all subjects in a population" (Lee & Foley, 1986).

In an attempt to confirm the Lee & Foley results, Waldman & Avolio (1989) found starkly dissimilar results following a study of 24,219 General Aptitude Test Battery (GATB) observations. They found that "the validity of the tests used in this study did not appreciably vary at different points along the test score range" (Waldman & Avolio, 1989). Specifically, "the slope of the regression and standard error of estimate did not significantly vary across the predictor score range" (Waldman & Avolio, 1989). Waldman & Avolio (1989) question Lee & Foley's conclusions due to the criterion selected by Lee & Foley. "Their (Lee & Foley) study did not address whether scores at different range levels differentially predicted performance on the job" (Waldman & Avolio, 1989).

2.1.2 Current Pilot Candidate Selection Method

The PCSM program was initiated in 1985 from an AETC/CC Program Guidance Letter and became operational in 1993 (Ness, 1996 and Carretta & Ree, 2000a). The current PCSM model is a weighted composite of seven items from three sources. First is the AFOQT Pilot composite, which is comprised of 8 of the 16 AFOQT sub-tests. See Table 3 for the specific sub-tests included in the AFOQT Pilot composite. Second, the BAT provides five scores used in the PCSM model, which are discussed in section 2.3.2.

Finally, a self-reported number of FAA flying hours is included. The weights for the seven PCSM inputs are based on a linear regression, which is then transformed by applying a discrete approximation of a sigmoidal function to the resulting linear combination. This particular sigmoidal function's origin is unknown by the author. PCSM scores range from 1-99. Figure 1 presents the authors understanding of the PCSM model architecture.

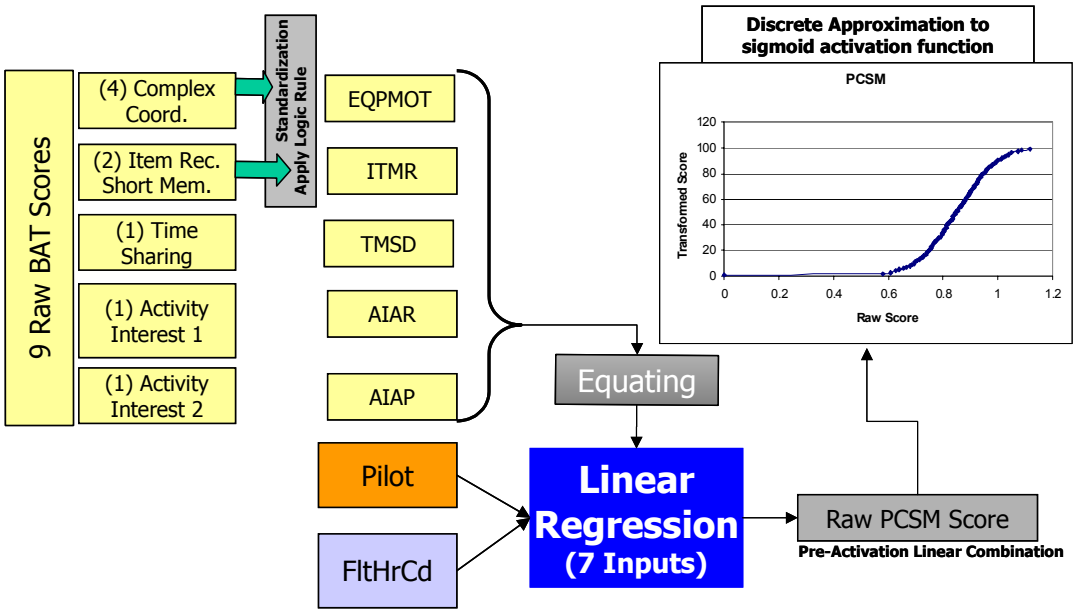


Figure 1. PCSM Model Representation

The EQPMOT score is calculated by standardization of its four input scores and averaged so that each input is equally weighted. This average is multiplied by –1.0 to account for the fact that the 4 inputs are pursuit tracking errors scores, hence larger scores imply worse performance. Item recognition reaction time is assigned a maximal score of 2,500 milliseconds if the percentage correct on the item recognition test is less than 75%. Finally, the five BAT inputs are transformed by way of equating tables.

Equating is necessary due to differences in equipment configuration of the experimental and operational forms of the BAT test. An analysis was done "to determine differences and where differences existed to equate the experimental and operational BAT tests" (Carretta & Ree, 1993). Equating is required so that operational BAT scores can be used in the current PCSM model, which was developed on data from the experimental BAT test configuration (Carretta & Ree, 1993). BAT scores have not been re-normed to the operational test. Therefore, the current PCSM model is still implementing equating tables, based on the equating study performed by Carretta & Ree (1993), to transform raw operational BAT scores prior to applying the regression weights.

"The validity of PCSM has been shown to come mostly from the measurement of cognitive ability (*g*), psychomotor ability, pilot job knowledge, and flying experience" (Carretta, 2000b). Generally, validity research for PCSM has focused on finding predictors that provide incremental validity to the AFOQT Pilot composite. The Pilot composite is highly *g* loaded. Flying experience validity has been found to decrease as a function of training phase (Carretta & Ree, 1995). Thus, flying experience is most predictive of success early in pilot training.

This has prompted current interest at AETC in the effects of updating the PCSM model to baseline all applicants at a minimum of 50 hours of flying experience. Currently each selected applicant has the opportunity to receive 50 hours of flight training to earn a private pilot license prior to entering UPT, but the current PCSM model uses the applicants FAA flight hours reported at the time of application. No such baseline is implemented in this research. The data for this research is based on each

applicant's actual self-reported FAA flying hours at the time the PCSM score was calculated.

2.2 Pitfalls in Ability Research and Pilot Selection

Carretta & Ree (2000a) detail 9 commonly overlooked methodological issues in ability research. They are (1) range restriction, (2) unreliability, (3) dichotomization of the criterion, (4) subgroup effects, (5) weighting of variables, (6) misunderstanding constructs, (7) misinterpreting factor analytic results, (8) lack of statistical power, and (9) failure to cross-validate. Each is discussed separately in this section. For this research, range restriction is perhaps the most critical of the methodological issues reported by Carretta & Ree (2000a). Significant attention is given to range restriction in section 2.2.1. The other pitfalls are addressed to a lesser extent.

2.2.1 Range Restriction

The Pearson correlation, "*r* is a consistent and efficient estimate of population rho only under the condition that *r* is obtained from a random sample of the population" (Duan, et al., 1997). Range restriction is a term used when the population correlation between a predictor and criterion is underestimated in a pre-selected subset of the population. After reviewing 700 criterion- related validity studies, Linn et al. (1981) state, "Procedures for correcting correlations for range restriction are desperately needed in highly selective situations." A graphical presentation of the effects of range restriction is provided by Sackett & Yang (2000). Burnham, Paulson, Andrews (1950) and Bryant (1972) present numerical examples of the correction procedure.

Several formulas exist for correcting range-restricted correlations. These are generally attributed to Pearson (1903) in the univariate case and Lawley (1943) in the multivariate case. Although Pearson (1903) is credited with developing the first correction formulas for range restriction for three univariate cases, (Ree et al., 1994) Thorndike (1949) popularized Pearson's work by describing the three "cases" in which to apply Pearson's formulas correctly. The three cases "require hard-to-obtain estimates of population variances, which account for their infrequent use" (Ree et al., 1994). Several authors present the univariate correction formulas (Linn et al., 1981, Ree et al., 1994, Sackett & Wade, 1983, and Sackett & Yang, 2000).

Lawley (1943) is usually credited with developing a general multivariate correction formula that allows for a scenario involving selection on multiple predictors (Ree et al., 1994); however, it is less commonly noted that Lawley extended work by Aitken (1934) (Sackett & Yang, 2000). The multivariate correction performs poorly for small samples, but has been shown to be more accurate (robust) for large samples even in the presence of assumption violations (Held, 1996).

The general assumptions underlying the application of the correction formulas provided below are taken from Duan et al. (1997), although they are commonly cited in the literature. The Lawley (1943) multivariate correction relaxes the normality assumption, which makes it appealing to researchers.

1. Linearity of regression of the criterion on the predictor
2. Homoscedasticity of the criterion error variance for all values of the predictor
3. Bivariate normality among the predictors

Lawley's multivariate correction formula differentiates between selection variables and incidental variables. Selection variables are those variables for which selection is based and are available for an unrestricted population (AFOQT scores, BAT scores, & other available predictors). Incidental variables are measured only in the restricted sample (i.e. UPT criterion). Ree et al. (1994) present and develop matrix equations for the multivariate correction formula. Sackett &Yang (2000) state that the multivariate correction should be used in cases where simultaneous or sequential selection is made on multiple variables. This is the case with the USAF UPT selection process; however, the criteria is either different across selection sources or is not specifically defined.

Carretta & Ree, who have been involved in a significant portion of PCSM research, opt for Lawley's (1943) multivariate correction in their work on PCSM. Lawley's (1943) multivariate correction is popular due to its relaxed set of assumptions. Where corrected correlations are used in this research, both the uncorrected and corrected correlations are reported. Furthermore, in keeping with past research by Carretta, Ree, & others who have done significant PCSM research, Lawley's (1943) multivariate correction technique is used in this research. A Windows® based application developed by Johnson & Ree (1994) named *RANGEJ* makes calculating the Lawley (1943) multivariate correction straightforward. The *RANGEJ* application is used in this research.

Correction for range restriction is suggested if the sample does not accurately represent the population for which prediction is needed. The need for range restriction corrections is common in ability research because information about the applicant

population is often lost or unavailable. For example, it is impossible to collect criterion scores for those not selected. The goal of this research is to create a model using predictor and criterion data on a pre-selected sample of those who previously attended UPT and use that model to predict the success of future pilot training applicants prior to selection for UPT. Since UPT selection is not done randomly, range restriction occurs.

Range restriction can occur as a result of many different selection scenarios. The two most common are direct and indirect (incidental) selection from the population. Under direct selection, applicants are selected based strictly on a cut-off score for a predictor. Under indirect selection, applicants are selected based on a cut-off score of a predictor or some combination of predictors not included in the available set. This causes range restriction in the predictor set, which is related to the magnitude of the correlation between the predictor on which selection occurred and the predictors in the predictor set. Range restriction can also occur when the selection process is unknown or based on immeasurable predictors.

2.2.1.1 Robustness and Accuracy of Corrections

Correction accuracy is a function of the degree to which the underlying data assumptions have been met, direct vs. indirect (incidental) selection, and the correction formula (univariate or multivariate) (Sackett & Yang, 2000). Historically, selecting a correction formula has been based on whether the data is univariate or multivariate alone. Sackett & Yang (2000) state "Thorndike correction formulas for range restriction can be shown to be special cases of a multivariate correction formula developed by Aitken (1934) and extended by Lawley (1943)."

Sackett & Yang (2000) discuss multiple scenarios under which range restriction commonly occurs and their effects on the restricted correlation and regression coefficients. They investigate a total of 11 different range-restriction scenarios based on combinations of three facets. The three facets are: (a) the variable on which selection occurs, (b) whether unrestricted variances for the relevant variables are known, and (c) whether a 3rd variable (indirect selection), if involved, is measured or unmeasured (Sackett & Yang, 2000). This significantly expands the menu of correction formulas available to the researcher.

Sackett & Yang (2000) found that "the Aitken-Lawley multivariate correction formula consistently reproduced population correlations closely, with precision decreasing as the sample size and the selection ratio decreased." Selection ratio is the proportion of applicants selected from the population. A small selection ratio could imply a highly selective process. Sackett & Yang's results confirm previous results by Lee et al. (1982) and Greener & Osburn (1980); however, the "closeness" of the corrected correlation was shown to degrade rapidly as selection ratio decreased in Greener & Osburn (1980).

Sackett & Yang's (2000) last case is most germane to this research in lieu of the differences in selection processes across selection sources found by Weeks (1998). The Weeks (1998) study is discussed in section 2.3.3. Sackett & Yang (2000) endorse more caution in using correction processes when the "range restriction processes are not fully understood, such as those in which unmeasured variables play a large role." Given the inaccuracies of correction formulas when applied to the incorrect range restriction scenarios, the researcher must take caution in selecting the appropriate correction formula

(Sackett & Yang, 2000). Such is the case with the selection process for UPT since the AFA and ROTC account for the majority of pilot candidates selected across the entire Air Force (Weeks, 1998). Weeks (1998) cautions researchers considering corrections for range restriction related to this research.

The results discussed to this point and those discussed in the next several sub-sections suggest that the range restriction does occur as a result of UPT selection; however, simply meeting the theoretical criteria known to induce range restriction does not guarantee corrected correlations are appreciably more desirable. Despite the generally positive results obtained through correcting correlations for range restriction, the true unrestricted population correlation is known in most studies on the accuracy and/or robustness of correlations corrected for range restriction. Unfortunately, unselected pilot applicants never go to UPT; therefore, no information on their criterion scores is available. Therefore, the accuracy of such corrections cannot be estimated.

Lord & Novick's (1968) text, *Statistical Theories of Mental Test Scores*" is often cited in literature related to the range restriction problem. "Lord & Novick (1968) indicated that there is a tendency for test score data to violate both assumptions (linearity & homoscedasticity) ... at both extremes of the distribution and expressed serious reservations regarding the probable accuracy of the corrections under conditions of extreme selection" (Greener & Osburn, 1980). To that end, Greener & Osburn (1979, 1980) specifically address the accuracy & robustness of corrected correlations.

Greener & Osburn (1979) performed an empirical study of the corrections in the case of direct selection. Greener & Osburn (1980) studied accuracy in the case of direct selection in the presence of differing degrees of either heteroscedasticity or non-linearity

of regression. Greener & Osburn (1979) also found the correction "was very sensitive to moderate departures from linearity (of the regression) but was quite robust in the face of rather substantial departures from homoscedasticity" (Greener & Osburn, 1979).

Greener & Osburn (1979) found that the accuracy of the correction is a function of the magnitude of unrestricted correlations in the population. Noting a difference in results, Duan et al. (1997) makes the opposite statement about the relationship between correction accuracy and magnitude of the unrestricted correlation. In Greener & Osburn (1979, 1980), the corrected sample correlation typically was more accurate than the uncorrected sample correlation when the unrestricted population correlation, rho, was moderate. However, Greener & Osburn (1979) found that the corrected correlation was no more accurate than the uncorrected one when the population rho was small (.10 to .25)" (Duan et al., 1997). This case is likely to hold for many of the predictors available in this research. Duan et al. (1997) point out that the discrepancy in results could be related to the fact that Greener & Osburn (1979) used empirical distributions, which may have violated the linearity and/or homoscedasticity assumptions. Greener & Osburn (1980) further studied corrections under assumption violations.

Duan et al. (1997) investigated the accuracy of several methods of estimating standard error of correlations corrected for Thordike's Case 2 range restriction via Monte Carlo simulation. Duan et al. (1997) found four main results, which follow. First, in every case investigated ($5 < N < 50$) "the correlation coefficient corrected for range restriction always was a more accurate approximation of the population rho than the correlation calculated from the restricted sample." This is consistent with Lin et al (1981) and Mendoza (1991). Second, Duan's et al. (1997) results suggest, "when the

selection ratio is very small the corrected Pearson correlation is not accurate in estimating population rho." Nevertheless, corrected estimates are still overly conservative in this case. Third, for a given range restricted rho and selection ratio, the corrected correlation became more accurate as sample size increased. This result agrees with Bobko's (1983) findings. The largest sample size in the Duan et al. study was N = 50. Finally, "the accuracy of *Rc* (corrected correlation) has no apparent relation to population rho (Duan et al., 1997).

Extending their research, Greener & Osburn (1980) simulated 9 bivariate distributions with correlations ranging from 0.10 to 0.90. From each one of these distributions, 7 samples (N = 40,000) were obtained for 5 different replications. The 7 samples included a bivariate normal distribution with equal means and variances and two each of three distributions, which violate one of the assumptions to differing degrees. The first type violated linearity of regression via a sigmoidal shape. The second type violated homoscedasticity with an increasing trend in variance (fan shaped). The last type, violated homoscedasticity by displayed an increasing, then decreasing variance trend (football shaped). Greener & Osburn (1980) simulated varying degrees of range restriction by truncating the lower portion of the sample with increasing proportions.

Greener & Osburn (1980) found that "violation of one or both of these assumptions can lead to serious errors in estimating the unrestricted population correlation." This suggests the need to verify assumptions whenever the correction formulas are used. Such has not been the case in PCSM related studies reviewed in this research.

Greener & Osburn (1980) found that correction accuracy is a function of both selection ratio (in a strict truncation fashion) and unrestricted population correlation. As selection ratio decreases and/or population correlation increases, the accuracy of the correction decreases. The correction accuracy for the sigmoid shaped samples (non-linear in regression) did not tend to be a function of the population correlation (Greener & Osburn, 1980). With one exception, the corrected correlations tended to be negatively biased across the Greener & Osburn (1980) study; however, in terms of percentage of error reported by Greener & Osburn the errors tended to be quite large. Lee et al. (1982) comment on the "gross underestimation" they witnessed. Over correction was seen in only in the football shaped samples when 60% or more of the sample was truncated from the lower end. Over correction tended to be less than the underestimate of the uncorrected correlations for all but the most severe truncation (80% to 90%). Greener & Osburn (1980) report the following results.

1. No tendency in the bivariate normal distributions to underestimate or overestimate population correlations regardless of the magnitude of population correlation and degree of truncation.

2. The corrected correlations for the sigmoid (non-linear) distributions were not affected by the magnitude of the population correlation, but were increasingly underestimated as the degree of truncation increased.

3. The corrected correlations for the fan-shaped (monotonically increasing heteroscedasticity) distributions increasingly underestimated the population correlation as a function of both the magnitude of the population correlation and the degree of truncation.

It appears the results of Greener & Osburn (1980) do not fully support the conclusions of Greener & Osburn (1979). Greener & Osburn (1979) noted the robustness of the correction "in the face of rather substantial departures from homoscedasticity."

For the heteroscedastic samples, the negative bias in the corrected correlations was as large or larger than those found in the non-linear samples. Although, all corrections generally tended to be negatively biased across the range of population correlations, the instances of over corrections found for the football shaped samples were smaller in magnitude than the bias in the uncorrected estimates for all but the most sever range restriction.

Lee et al. (1982) found a slight overestimation in corrected correlations. The correction changed the correlation from 0.59 to 0.75 when the true population correlation was 0.68. This occurred while implementing both triple and double correction procedures to correct for both range restriction and unreliability. Lee et al. (1982) point out that their use of split-half reliability may not be an optimal estimate for criterion reliability in the process of the correction. Thus, using the split-half reliability estimates for predictor and criterion to correct for unreliability in the multi-correction procedure could be the cause of inconsistent results compared to Greener & Osburn (1980).

Lee et al. did confirm that "uncorrected coefficients grossly underestimate the true validity ... and that the magnitude of the underestimate is inversely related to the selection ratio" (Lee et al., 1982). This agrees with Greener & Osburn (1980). Lee et al. (1982) used selection ratios that ranged from 0.50 to 0.10. Although in some cases, corrected correlations overestimated population correlations, their results for the corrected correlations confirm this inverse relationship as the bias in corrected correlations decreased from about 9% overestimation to more than 10% underestimation when selection ratio decreased from 0.20 to 0.10.

Gross & Fleischman (1983) investigated simultaneous violation of distributional and selection assumptions using actual test score data. In previous studies on the effects of distributional violations, selection was assumed to be explicit (i.e. performed on a single variable). Primarily, Gross & Fleischman (1983) make two conclusions.

> First, it is unreasonable to assume that the correction formula can exactly reproduce or even closely approximate the total group correlation when neither the underlying distribution nor the selection assumptions are violated. At best reasonably small percentage errors in the range of 15% to 20% can be assured only when the degree of selection is quite modest.

Thus, if the unrestricted correlation estimate requires a high degree of accuracy, "the correction formula will be inadequate, especially as the proportion of missing y (criterion) scores increases." Such accuracy is required when selecting predictors during model development. Inaccuracy in the corrected correlations may cause more problems than the uncorrected correlations themselves. The corrections are not robust to violations of both assumptions and errors found were reasonably small, "only for very modest degrees of selection" (Gross & Fleischman, 1983). "As the proportion selected (from the population) decreases, the accuracy of the formula deteriorates" (Gross & Fleischman, 1983).

Secondly, Gross & Fleischman (1983) found accuracy to be highly dependent on the distribution form underlying the data when both assumptions are violated. Specifically, distributional forms "where the regression curve is exponential in form, and the variance of y (criterion) is a decreasing function of X1 (the predictor)" results in a substantial overestimate (Gross & Fleischman, 1983). Furthermore, Gross & Fleischman

(1983) state that in some cases, the uncorrected correlation is a better estimate of the unrestricted population correlation.

2.2.1.2 Conditions Resulting In Conservative Corrections

Linn et al. (1981) examined over 700 criterion-related validity studies in an attempt to investigate:

1. The relationship between the standard deviation of the predictor and the magnitude of the predictive validity (correlation).
2. Estimate the effect of corrections for range restriction assuming explicit (direct) and incidental (indirect) range restriction.

A strong positive relationship was found between the standard deviation of the predictor and the magnitude of the predictive validity (Linn et al., 1981). As the standard deviation in the predictor decreases as a result of range restriction, the validity of the predictor also decreased. Linn et al. (1981) found that the corrections reduced the strength of this relationship, thus they are considered better than the uncorrected validities, but "still apt to provide a conservative estimate." This agrees with Greener & Osburn's (1979,1980) findings that corrected correlations tend to increasingly underestimate population correlations as a function of selection ratio, despite violations of the linearity or homoscedasticity assumption.

The goal of studying these relationships was to ultimately investigate "the combined effects on corrections of violations of assumptions (linearity in regression and homoscedasticity) and selection on an unspecified variable (direct or indirect selection)" (Linn et al., 1981). Linn et al. (1981) cite Brewer & Hills (1969) for finding that "inaccuracy of the corrections increased with increasing skewness and increasing

selectivity (decreasing selection ratio)." Linn et al (1981) cite earlier work by Linn (1968) that suggest corrections are negatively biased under the indirect selection scenario when there exist strong correlation between the true selection variable and the available predictor variable. Strong correlations have not been shown in the case of PCSM. This is evidenced by the fact that USAF selection processes have resulted in selecting applicants whose combined PCSM scores from 1993-2001 (N=18,927) that are fairly uniformly distributed across the PCSM score range. In fact, more applicants have been selected with low PCSM scores than high scores. Figure 61 in chapter 5 presents a histogram of all valid PCSM scores in the data provided for this research.

Table 2 summarizes Sackett & Yang's (2000) findings for several range restriction scenarios a when positive correlation exists between a predictor and the criterion in the population.

Table 2. Sackett & Yang (2000) Study Findings

Range Restriction Scenario	Unrestricted Correlation	Regression Coefficient
Direct on predictor	underestimated	unaffected
Direct on criterion	underestimated	affected
Direct on extremes	overestimated	unaffected
Indirect on a 3rd predictor	underestimated	affected

The following are three considerations for correcting correlations for range restriction suggested by Sackett & Yang (2000).

1. Sampling error of corrected correlations
2. Robustness of correction formulas against violations of linearity and homoscedasticity

3. Using maximum likelihood methods as an alternative for dealing with missing data.

The second consideration, robustness to violations of assumptions, was discussed in section 2.2.1.1. Furthermore, conditions that result in conservative corrections were discussed earlier in this section. The first consideration, sampling error, is discussed in the next section. The reader can go to Sackett & Yang (2000) for a more information on maximum likelihood methods.

2.2.1.3 Sampling Error of Corrected Correlations

Gross & Kagen (1983) stated that very little is known about the standard error and sampling distribution of corrected correlations. Mendoza et al. (1991) and Greener & Osburn (1980) both cite Forsyth (1971) stating that the traditional calculation for standard error of the Pearson product moment correlation coefficient is not appropriate for confidence intervals about corrected correlations. More recently, Ree et al. (1994) state that corrected correlations do not have a known sampling distribution and standard error; therefore, statistical significance tests are theoretically not possible with corrected correlations. Other methods of establishing confidence intervals around corrected correlations had not been investigated when Forsyth (1971) was published. Since then, Bobko (1983) used Talyor series expansion to develop standard error estimates, thus making confidence intervals possible for correlations corrected for both range restriction and attenuation.

Bobko (1983) showed that although standard error increases for doubly corrected correlations (corrected for both range restriction and unreliability), "the proportionate increase in standard error is less than the gain in magnitude in correlation, thus,

confidence intervals for double corrected correlations, though wider, are narrower than one would expect from the increase in magnitude of the point estimates." Bobko's development depends on large sample sizes and assumes the ratio of the "applicant to restricted variance is fixed and known" (Mendoza et al., 1991). The criterion variance in the unrestricted applicant population is not known in the case of UPT applicants. Furthermore, due to changes in pilot production quotas reported by Weeks (1998), this unknown variance is most likely not constant.

Bobko (1983) also demonstrated that estimates of double corrected correlations are negatively biased, which is "in direct contrast to the empirical conclusion of Lee et al. (1982)." This follows from Bobko's interpretation of the individual impacts of three multiplicative terms in the bias equation he presents. Bias increases as criterion reliability decreases for the first term of the equation. Bias increases as the selection ratio decreases for the second term. Bias approaches zero as sample size (n) becomes sufficiently large for the last term. The multiplicative effect of these terms is that the overall correction is expected to have a negative bias. The reader is directed to Bobko (1983) for a proof that the double corrected correlation bias is negative.

Mendoza et al. (1991) demonstrated a Bootstrap method for obtaining confidence intervals on the unrestricted population correlations from four simulated distributions. These distributions are normal, mixed, positively skewed, and negatively skewed. Mendoza et al. (1991) found positive results compared to other confidence interval building methods. "The corrected correlation coefficient yielded accurate bootstrap intervals over the four distributions" (Mendoza et al., 1991). Although population rho

(correlation) had little affect on confidence interval accuracy, small population rho (R = 0.1) coupled with small sample size affected stability (Mendoza et al., 1991).

Mendoza's (1991) bootstrap method for calculating a confidence interval around the unrestricted correlation requires only general assumptions, which makes it more applicable for selection studies with small to moderate sample sizes than Bobko's (1983) Taylor series based method. Mendoza et al. (1991) "found that the size of population rho (correlation) did not affect the accuracy of the confidence interval."

2.2.1.4 Multiple Correction Procedure

Although Bobko (1983) & Greener & Osburn (1980) do not substantiate Lee's et al. (1982) results, Lee et al. (1982) provided a discussion and comparison of two "double correction" methods, which correct for both range restriction and unreliability of the criterion. Lee et al. (1982) cite Schmidt et al (1976) and Nunnally (1978) for suggesting "it would be inappropriate to correct for unreliability in both the criterion and the predictor." Bobko (1983) provide the following conclusions about double corrected correlations under the assumptions of underlying bivariate normality and adequate sample size.

1. Bias is inversely proportional to sample size.
2. Standard error is inversely proportional to the square root of the sample size.
3. Overall bias is negative.

The first two properties listed here are also true of uncorrected correlations (Bobko, 1983). The expressions for bias and the variance of the corrected correlation are inversely proportional to the square root of the criterion reliability and the selection ratio,

thus the corrected correlation is less accurate with smaller reliability and/or selection ratio (Bobko, 1983). Confidence intervals can be reported for corrected correlations, which are narrower than one would expect from the increase in the point estimates (Bobko, 1983). Bobko & Rieck (1980) "have indicated that Taylor series approximations of single corrected correlations are precise if *n* is greater than 100." In a study on the validity of AFOQT test scores, Carretta & Ree (1993b) found that range restricted correlations further corrected for unreliability resulted in "trivial" changes. Therefore, only the correction for range restriction is employed in this research.

2.2.1.5 Sign Changes As A Result of Corrections

Ree et al. (1994) discuss sign changes that they have witnessed in corrected correlations using Lawley's multivariate formula. They explain these sign changes by close examination of the Lawley (1943) multivariate correction, which assumes selection is based on *p* predictor scores available in the unrestricted population. The correction involves estimation of a variance-covariance matrix that is divided into a *2 x 2* matrix of variance-covariance sub-matrices. In the development, Ree et al. (1994) show that unknown variance-covariance sub-matrices can be estimated from distributional information estimated for the unrestricted population. See Ree et al. (1994) for the full development of the Lawley (1943) multivariate correction.

Held (1996) provided further explanations for the sign changes discussed by Ree et al. (1994). Held states that under the assumption that all selector variables and the criterion are positively correlated, a negative to positive sign change in the corrected correlation is a "function of the inter-correlations of the selectors and criterion in the restricted data set, and cannot be viewed as an abnormal outcome." On the other hand, a

positive to negative sign change "may be a function of small and/or inadequate data set, and should be viewed as an unrealistic outcome" (Held, 1996).

2.2.1.6 Arguments Against Corrected Correlations

Damos (1996) is among the few authors who flatly reject correcting for range-restricted correlations. Although Damos points out some pertinent issues that are likely to combine with the range restriction phenomenon in causing correlation shrinkage in a sample, she does not provide evidence for her claim that "range restriction is a red herring." She seems to base her opinion on the fact that many corrected correlations are appreciably greater than the uncorrected correlations in the range-restricted sample. Although corrections resulting in overestimates of the population correlation are uncommon, they have been discussed in previous sections of this research. Damos (1996) does not offer any of those results as evidence of her case.

With that said, Damos points out several causes of low correlations in samples that are germane to the PCSM data available for this research. Carretta, Ree, & others who have performed PCSM research have not addressed several of these. The following examples cause greater concern for a researcher's ability to find variables that are truly capable of effectively predicting pilot training success. These include (1) sudden changes in pilot production quotas, (2) changes in pass/fail criteria and/or selection criteria as a result of changes in pilot production quotas, and (3) severe dichotomization of the criteria (Damos, 1996).

2.2.2 Reliability of Predictor Scores

The extent to which a predictor score is unreliable affects the magnitude of the criterion/predictor correlation. Reliability quantifies measurement error and ranges from zero to one. Reliability of a predictor score defined as "the ratio of true variability to total variability" (Ree & Carretta, 2002). Ree & Carretta (2002) state that false conclusions and interpretations can result from ignoring the consequences of using predictors with less than perfect reliability.

Reliability plays a role in many common statistical techniques, which are discussed in Ree & Carretta (2002). Any observed predictor score can be thought of as comprising the true score and uncorrelated error component. The variance of the observed score can be represented as the sum of the true and error variances as

$$\sigma^2_{obs} = \sigma^2_{true} + \sigma^2_{error}, \qquad \textbf{(2)}$$

where Ree & Carretta (2002) use the following definition of reliability when test/retest data are available:

$$r_{xx'} = \frac{\sigma^2_{true}}{\sigma^2_{obs}} \qquad \textbf{(3)}$$

Although the mean of a predictor or criterion score is not affected by unreliability, lower reliability implies that measurement error is higher, which causes increased observed variance (Ree & Carretta, 2002). Since reliability affects observed score variance, it also affects hypothesis tests and confidence intervals involving a measure of

standard error, which is based on standard deviation of the measure. The result is that tests are less sensitive and confidence intervals are wider, hence statistical power can be expected to decrease (Ree & Carretta, 2002). Ree & Carretta (2002) feel that decreased statistical power and wider confidence intervals could lead to misinterpretations of what constructs are being measured.

Unreliability also causes concerns for accuracy of correlations. "The magnitude of the correlation between variables is limited by their reliabilities" (Carretta & Ree, 2000a). "According to classical test theory, the upper bound on the validity (correlation coefficient) is the square root of the reliability" (Stanley, 1971). Lower reliabilities of two different measures of the same construct will cause the correlation between the scores from those measures to decrease (Ree & Carretta, 2002). A well-known formula that demonstrates the correlation between two variables as a function of the reliabilities of the two variables is cited by Ree & Carretta (2002) and is presented in Equation 4.

$$r_{xy} = r_C \left(\sqrt{r_{xx'}} \sqrt{r_{yy'}} \right) \quad \textbf{(4)}$$

Correlations can be corrected for unreliability by solving the above equation for r_c, the true correlation, which results in Equation 5 (Ree & Carretta, 2002).

$$r_C = \frac{r_{xy}}{\sqrt{r_{xx'}} \sqrt{r_{yy'}}} \quad \textbf{(5)}$$

Carretta & Ree (1993b) corrected for unreliability in conjunction with correcting for range restriction in several studies. Under hypothetical perfect reliability among three measures of a single construct and true correlation of 1.0 for any pair, partial correlation would be zero (Ree & Carretta, 2002). Ree & Carretta (2002) show that unreliability of 0.8 causes the partial correlation to increase to a moderate 0.44.

For multiple regression, "the effect of the reliability is a function of reliability magnitudes and the true score correlations among the predictors" (Ree & Carretta, 2002). Again, there is no biasing effect on the regression coefficients, but "the effect of the unreliability of the variable being partialed out has a substantial effect on the partial regression coefficient of the other variable (in the two predictor case)" (Ree & Carretta, 2002). Hence, "the uncorrected regression weights are not dependable indicators of the importance of the independent variables" (Ree & Carretta, 2002).

Ree & Carretta (2002) state that standard error of estimate increases as a result of unreliability by

$$\Delta SE_y = \sigma_y \left[\sqrt{1 - \left(r^2_{xy} r_{xx'} \right)} \left(\sqrt{1 - r^2_{xy}} \right) \right], \qquad \textbf{(6)}$$

while the standard deviation of the predicted criterion decreases by

$$\Delta \sigma_y = \left(1 - \sqrt{1 - r^2_{xx'}} \right) * \sigma_y, \qquad \textbf{(7)}$$

and validity coefficients are reduced as a result of unreliability by

$$\Delta r_{xy} = \left(\sqrt{1 - r^2_{xx'}}\right) * r_{xy} \quad \textbf{(8)}$$

In factor analysis, unreliability reduces factor loadings (Carretta & Ree, 2000a). This leads to erroneous conclusions about construct interpretation (Carretta & Ree, 2000a). As reliability improves, factor loadings can be interpreted more directly. Carretta & Ree (2000a) state that factor loadings can be corrected for unreliability. To do so "the underestimation can be corrected by dividing the factor loading by the reliability" (Carretta & Ree, 2000a). No test/retest data is available in this research; however, variable communalities can be used as lower bound estimates of the reliabilities (Ree & Carretta, 2002).

These issues are considered as part of the analysis, but the corrections discussed above are not applied. This decision is based on the lack of significant results found by Carretta (1994) and Carretta & Ree (1993b) after correcting for unreliability. In terms of validating the current PCSM model via stepwise multiple linear regression, the significant variables must be considered in light of these artifacts. However, the actual regression coefficients are not the focus of the analysis. In terms of updating the regression weights for the current PCSM model, no variable selection is performed. Further, no information available suggests that the current PCSM regression weights have been corrected for unreliability.

Ree & Carretta (2002) also suggest the use of "latent variable analyses" such as confirmatory factor analysis to "eliminate or substantially reduce the unreliability of the variables is a another worthwhile approach." In this research, factor analysis is used in this confirmatory sense as part of the PCSM validation. Despite any underlying affects

caused by predictor unreliability, varimax rotated factor loadings provided for a very straightforward construct interpretation relative to the current PCSM inputs.

2.2.3 Dichotomization of Criteria

In general, Cohen (1993) prefers the use of continuous criteria whenever possible. Cohen (1983) cites Cohen & Cohen (1983) for arguing against needless dichotomization of a criterion because "it results in underestimating effect sizes and reducing the power of statistical hypothesis test." Dichotomization has its roots in what is known as "broad" or "coarse" grouping of continuous variables as a method of simplifying statistical calculations (Cohen, 1983). In Cohen's view, the advent of the modern computer makes such data simplification unnecessary.

Cohen (1983) specifically discusses the case of a bivariate normal population where variable X predicts criterion Y. Dichotomization of the predictor variable at the mean, reduces the proportion of criterion variance accounted for by the dichotomized predictor variable to 64% of that accounted for by the undichotomized predictor variable (Cohen, 1983). The situation worsens as the point of dichotomization moves away from the mean. This results in reduced product-moment correlation and smaller test values, which obviously affect associated statistical tests (Cohen, 1983). With dichotomization being performed at or near the mean, large sample sizes can still detect significance with test values at approximately three-fifths to three-quarters as large as they should be without dichotomization (Cohen, 1983). However, sample size in no substitute for severe dichotomization.

In this research, the criterion is dichotomized. Unfortunately, there is no convenient way to create a continuous variable for the binary pass/fail criterion available

in this research. Carretta & Ree (1993) attempted to overcome this problem by creating a rank index based on actual UPT performance averages for those identified as successful. For eliminees, a ranking index based on the total number of flying hours completed prior to failure was fabricated such that the highest failure had a ranking index lower than the worst graduate. This provided a continuous looking criterion to use in a linear regression. Unfortunately, the number of flying hours completed is not available in this research; therefore, there exist no method for ranking failures. However, Carretta & Ree (1993b) found a correlation of 0.98 between the predicted outcomes for the binary pass/fail criterion and the ranking index criterion. In essence, the creation of a more continuous criterion had little effect on the actual predictions when the criterion was dichotomized on a data set. The data used in this research is similar in content to the data used in Carretta & Ree (1993b).

2.2.4 Subgroup Effects

Group effects may occur when multiple groups (i.e. sex or race) are represented within a sample. Validation of predictors is often based on correlation and regression. It is possible that the correlations between the individual groups and the criterion differ from the correlation between the combined group and the criterion. When this occurs, the regression equation based on the combined group can lead to inaccurate conclusions regarding the validity of the predictor for all groups represented in the sample (Ree & Carretta, 1999). Ree & Carretta (1999) discuss possible scenarios and the resulting effect on the regression equation for the combined group.

Ree & Carretta (1999) provide an example of "Simpson's paradox" in which each sub-group has zero correlation with the criterion, but group mean differences on the

criterion produces a positive correlation between the combined group and the criterion. When two groups exist in a sample, the authors suggest a hierarchical linear models approach. "Hierarchical analysis consists of a series of linear regression models that are tested to determine contributions of independent variables to prediction of the dependent variable" (Ree & Carretta, 1999). The series of tests allows the researcher to systematically investigate differences in regression slopes and intercepts of the individual groups and combined group models to determine if a final model with only one slope and one intercept based on the combined group is appropriate (Ree & Carretta, 1999). To use a single model for prediction among the combined group is invalid when models for the subgroups result in significantly different parameters.

For situations involving more than two groups (i.e. Race), Ree & Carretta (1999) suggest using Within and Between Analysis (WABA) and cite Dansereau et al. (1984) for providing a detailed discussion of WABA. Race is not considered in the current PCSM model; therefore, race is not considered in this research.

Carretta & Ree (1995) and Carretta (1997a, 1997b) investigated both sex and race/ethnic group differences between USAF officer and USAF pilot applicants. Factorial invariance is a term used to describe whether selection instruments measure the same factors for all groups (Carretta & Ree, 2000a). In other words, factorial invariance implies that predictor scores load on the same factors in all subgroups. If factorial invariance does not hold, misinterpretation of the constructs being measured may occur because the set of predictors used to define an underlying construct is subgroup dependent (Carretta & Ree, 2000a).

In a study of a large sample of USAF officer applicants, Carretta & Ree (1995) examined the factor structure of the AFOQT. The study results showed "nearly identical structure of ability for sex and race/ethnic groups" (Carretta & Ree, 2000a). Since all USAF pilots are officers, these results also apply to the USAF pilot applicant population of interest in this research. In a study of USAF pilot applicants, Carretta (1997a) investigated the factor structure of the BAT in terms of sex. Recall that the BAT provides 5 of 7 inputs into the PCSM model. "Despite means score differences (among sex subgroups) on the tests, results indicated near identity of factor structure for men and women" (Carretta & Ree, 2000a).

Once factorial invariance is demonstrated, the researcher should test for differences in mean scores among the sub-groups. Carretta & Ree (2000a) cite several recent studies by Carretta (1997a, 1997b), in which mean score differences for the AFOQT and BAT among sex and race/ethnic groups where investigated for USAF officer applicants and pilot trainees. For the AFOQT composite scores, a significant difference in mean scores were found for groups in terms of sex (male vs. female) and race/ethnicity (Whites vs. Blacks vs. Hispanics). In both cases, the selection process is thought to be the cause of a reduction in the mean score differences in terms of standard deviation units ($\underline{d}$) by $\geq$ 50% (Carretta, 1997a). For the BAT, mean score differences were reported for sex subgroups only, hence Carretta (1997b) was unable to report race/ethnic results. Carretta & Ree (2000a) report that all mean score differences favored males and were statistically significant for the sample of USAF pilot applicants. No change in standard deviation units ($\underline{d}$) was reported for those selected for pilot training.

Differential validity occurs when tests do not measure the same constructs for different groups. Referring to the wealth of literature available, Carretta & Ree (2000a) state "the cumulative evidence overwhelmingly demonstrates that differential validity is almost nonexistent for cognitive tests," which accounts for the majority of PCSM input. Carretta's (1997a) results confirm this lack of differential validity for the AFOQT.

In summary, despite the presence of group mean score differences in terms of sex and race/ethnicity for USAF pilot trainees, factorial invariance holds and differential validity does not hold. Both situations are favorable in terms of using factor analysis to explore and understand potential predictive scores.

2.2.5 Weighting of Variables

Weighting of variables refers to the linear combination coefficients derived by some optimization procedure such as regression. "Two common weighting methods include unit weighting and criterion-based regression weighting" (Carretta & Ree, 2000a). Walters et al. (1993) cite criterion-based regression weighting as the norm in pilot selection (Carretta & Ree, 2000a). Ree et al. (1998) showed that unit weights produced "nearly identical rank orders of candidates when compared with other weighting schemes." Carretta & Ree (1998 & 2000a) argue for using unit weighting whenever top-down selection is used. According to Weeks' (1998) policy capturing study for four sources of UPT candidates, top-down selection is the primary method of selecting pilots. However, the situation is exacerbated by the fact that each selection source uses its own ranking system and PCSM is not the most significant factor for the ranking systems of the AFA and ROTC, who select the majority of all UPT candidates.

Although unit weighting has intuitive appeal for understanding the resulting composite, it provides for less intuitive means of comparing candidates, which is important in pilot selection. Pilot candidates near the "cut-off" are compared before the final selections are made. Neither a candidate's rank order or unit weighted composite provide a normative scale for which to make a comparison (Ree et al., 1998). Such a normative scale provides a means for making interpretations about the magnitude of the difference between two composite scores, which are important when comparing two applicants or when a data set spans multiple years. While PCSM's regression based weighting may be less intuitive than unit weights, the fact that its output has a probabilistic interpretation makes it ideal for comparing two candidates. On the other hand, "simple and unit weights are not influenced by outliers in the data and cannot lead to shrinkage on cross application (cross-validation)" (Ree et al., 1998), a benefit not enjoyed by other regression-based methods.

Ree et al. (1998) offer the conclusion that similarity in rank order between regression-based weights and unit weights could be explained by an average correlation (0.60) among the 10 Armed Services Vocational Aptitude Battery (ASVAB) measures used as predictors in that study of enlisted Air Force members. It is certainly true that similarity in regression weights across job families caused similar rank order results (Ree et al., 1998). To counter this problem, Ree et al. (1998) also employed random weights (1-9) in place of the regression weights. The results were surprisingly similar to those rank orders found using unit weights and regression weights.

These results demonstrate Wilks' Theorem, which explains the "mathematical inevitability of the ubiquitous finding that unit weighting produces a composite that is

very highly correlated with composites weighted by any other method" (Ree et al., 1998). Inspection of Wilks' approximation formula (Equation 9) for the expected correlation of two weighted linear composites reveals that the correlation is proportional to the average correlation among the predictor variables and the number of predictor variables. In the Ree et al. (1998) example, the average correlation (*r*) among the 10 ASVAB scores was 0.60. The expected correlation is also dependent on the magnitude of the squared coefficients of variation (CV) (in brackets) of the two populations being compared. In Ree's et al. (1999) study, the CV's were all near unity, thus essentially acted as a constant.

$$\overline{R} = 1 - \frac{1}{2\overline{r}K}\left[\frac{\sigma^2{}_v}{\mu^2{}_v} + \frac{\sigma^2{}_w}{\mu^2{}_w}\right], \qquad \textbf{(9)}$$

where the terms of order $1/K^2$ ... $1/K^n$ are dropped, (*r*) is the average correlation of the variables, K is the number of variables, *v* and *w* represent randomly drawn weight sets.

Despite the simplicity of this consequence, viewing PCSM scores as a pseudo probabilistic inference of success plays a large role in operational implementation for this research. The ability to interpret a PCSM score as a probability of success in pilot training must be maintained in order to meet the customer's operational requirement. Therefore, unit weighting will not be used in this research.

2.2.6 Misunderstanding Constructs

Constructs are abstractions of abilities that researchers seek to measure. Once a construct is identified, the researcher seeks to find measures that have construct validity. One example of a construct germane to this research is that of Officership. Weeks (1998)

found measures of Officership to be significant in the AFA and ROTC pilot selection processes; however, the validity of the Officership measure used at ROTC was not found to be statistically significant ($p > 0.05$). No significance test for the Cumulative Military Performance (CMP) measure used at the AFA was reported by Weeks (1998). Other constructs commonly used in military applications include intelligence, leadership ability, or situational awareness. Constructs are not observed directly, but are inferred (Carretta & Ree, 2000a). What a predictor appears to measure is not necessarily what it actually measures. Without studying construct validity formally, many researchers make erroneous assumptions about what is being measured by tests (Carretta & Ree, 2000a).

The factor structure and validity of the AFOQT and BAT tests has been formally investigated in several studies (Carretta, 1997, Carretta & Ree, 1994, 1995a, 1995b, 1996). Specifically, the AFOQT and BAT tend to be "*g*-loaded" on the general cognitive ability construct (Carretta, 1997, Carretta & Ree, 1995b, 1996, Ree & Carretta, 1995, Olea & Ree, 1994). "A long history of research findings has demonstrated *g* to be the most valid predictor of academic performance, job performance, and for numerous other human characteristics" (Carretta & Ree, 2000a). However, there are those who place less confidence in *g* as a predictor of ability (Bauer, 2003b).

In a study by Wheeler & Ree (1997), "results indicated that the validity of the BAT psychomotor tests comes from their measurement of a general psychomotor factor and *g*" (Carretta & Ree, 2000a). Olea & Ree (1994) studied the validity of general cognitive ability and special abilities (spatial or perceptual) to predict several pilot criteria including academic training performance and flying work samples. Olea & Ree (1994) used AFOQT scores to estimate the predictors. Again, *g* was found to be most significant

with specific abilities being incrementally valid to *g* (Olea & Ree, 1994). Ree & Carretta (1998) found very high correlation (near 1.0) between verbal and quantitative scores on the AFOQT and Scholastic Aptitude Test (SAT), respectively. This confirms that these AFOQT scores capture the same construct as the SAT composites, which are accepted measures of general cognitive ability, *g*.

Carretta & Ree (2000a) hypothesize that the incremental validity of specific abilities over the AFOQT pilot composite was due to specific aviation related job knowledge "rather than specific abilities such as spatial or perceptual ability." Carretta & Ree (2000a) summarize by stating, "research results point to *g* as the most important underlying construct in the prediction of pilot training success."

Jones & Ree (1998) found that "job ability differences did not moderate the relationship between the amount of *g* measured by a test and its score validity. This study was accomplished both across and within a range of job families requiring differential job skills (Jones & Ree, 1998). Dissimilar results do exist. Carretta & Ree (2000a) reported on a survey of several meta-analyses, which showed "measures of cognitive ability and personality were less valid." They concluded that these results could be expected because such measures are "mainstays in military pilot selection procedures, thus leading to restriction of range on these constructs." Carretta & Ree (2000a) cite Jensen (1998) for a complete presentation and discussion of general cognitive ability.

Carretta & Ree (1996) performed a confirmatory factor analysis (CFA), which "found that the AFOQT displays a hierarchical nature similar to other multiple aptitude test batteries." The higher-order factor found in this CFA was identified as general

cognitive ability (g), which accounted for 67% of the common variance with all 16 AFOQT sub-tests contributing to the measurement of *g* (Carretta & Ree, 1996). These 16 sub-tests are then used to create the 5 AFOQT composites discussed in Section 2.3.1.

As noted throughout this section, the factor structure of the AFOQT & BAT scores currently used in the PCSM model has been studied by Carretta (1997), Carretta & Ree (1995b, 2000a), and Ree, Carretta, & Earles (1999a). This research makes no attempt to further understand or define the factor structure or constructs measured by the AFOQT or BAT composites. Factor analysis is only used for the purpose of validating the those predictors currently in the PCSM model.

2.2.7 Misinterpretation of Factor Analytic Results

Factor analysis seeks to determine and explain unobservable sources of variation in a correlation matrix. Factor analysis is often used to identify latent constructs being measured by inspecting the factor loadings of each predictor across a predetermined number of factors extracted via factor analysis. This is normally done via eigenvalue analysis. By grouping predictors with factor loadings that meet some threshold, normally 0.50, it is possible to form an interpretation of what is being measured by each group of predictors. Carretta & Ree (2000a) warn that the standard practice of factor rotation can cause misinterpretation due to a phenomenon known as the disappearing first factor. Carretta & Ree (2000a) suggests foregoing rotation or using a residualized hierarchical solution, such as that used in a CFA of the AFOQT (Carretta & Ree, 1996).

Residualized hierarchical analysis is a method that uses factor analysis to identify a single higher-order factor, which accounts for the most common variance among the predictors. The common variance accounted for is generally referred to as communality

of the factor. The remaining proportion of the variance is considered to be unique and can then be attributed to lower-order factors. Carretta & Ree (2000a) cite Schmid, J., & Leiman, J.M. (1957) as a source of more detail on developing hierarchical factor solutions.

Not withstanding, rotation methods are well understood and routinely used in factor analysis as a method of redistributing factor loadings. Factor rotation seeks to maximize the factor loadings of each predictor on a single factor, while minimizing the factor loadings of that predictor on all other factors. Thus, interpretation of the underlying latent factors is made more apparent. However, there does exist some reason for concern. Rennie (1997) provides the following quote on the matter of the appropriateness of rotation from Pedhazur & Schmelkin's (1991, p. 611) textbook, "What might be viewed as a meaningful rotation from one theoretical perspective may not be considered meaningful, even utterly inappropriate, from another." Rennie (1997) cites Hetzel (in press) by stating, "with varimax rotation, there is a tendency for the principal factor to disappear because the factor variance is redistributed." This redistribution is exactly what researchers who employ rotation methods rely on to make clear a plausible factor interpretation. In terms of construct validation, the question for Rennie (1997) is not whether or not to rotate, but rather which rotation method best suites the researcher's current application. Rennie states, "rotation is used in almost all exploratory factor analysis studies."

In the present research, factor analysis is used in a confirmatory sense and as a data reduction technique. Rather than make interpretations of the constructs being measured by the AFOAT and BAT tests, these tests become the identifier for their

construct. Data reduction results from factor analysis are compared for both unrotated and varimax rotated factors. The factor analysis performed in this research is used to confirm that the inputs of the current PCSM model are still valid at the construct level. No attempt to duplicate CFA work discussed in this section is made. Therefore, no residualized hierarchical analysis is performed.

2.2.8 Lack of Statistical Power

Sackett & Wade (1983) demonstrated that statistical power is much better than Schmidt et al. (1976) suggest for the average size validity study (N = 68) when indirect range restriction occurs. "Under indirect range restriction, the average validity study (N = 68) has a 75% chance of detecting validity if validity exists when a one-tailed test can be used" (Sackett & Wade, 1983). Under direct range restriction, "their (Schmidt et al., 1976) tables are appropriate (Sackett & Wade, 1983). In this research, direct selection does not occur on any of the predictors available. In fact, the indirect nature of the range restriction that occurs during the selection of UPT candidates in not understood because each selection source follows a unique selection process. Therefore, the relaxed sample sizes reported by Sackett & Wade (1983) are applicable.

The required sample size is a function of required statistical power, criterion reliability, and selection ratio for a given combination of experimental predictor validity, interpredictor correlation, and operational predictor validity (Sackett & Wade, 1983). Selection ratio is the proportion of the unrestricted population selected. In their study involving common combinations of experimental and operational predictor validities and interpredictor correlation, Sackett & Wade (1983) found that the required sample size for

a required power of 0.90 ranged from only 32 to 312. Training set sample sizes used in this research range from 282 to 6,310.

2.2.9 Failure to Cross-Validate

The squared cross-validity coefficient, R_c^2, is often used to estimate the predictive power of a sample regression equation for use in future samples from the population. Kennedy (1988) defines R_c^2 as the "squared correlation of the actual criterion values with the predicted values from the sample equation for the population of interest." Under typical social applications with moderate sample size, significant loss of information, due to cross-validation, causes inflation of the validity estimates after empirical model selection (Kennedy, 1988). This effect is counter to the effects of range restriction, which causes the unrestricted correlation to be underestimated in the sample. The effect is less dramatic with larger samples (Kennedy, 1988), such as those available in this research. Kennedy also noted work by Hockings (1976), Rencher & Pun (1980), and Lerner & Games (1981) that supports this conclusion.

Loss of information for model development due to holding-out part of the sample for validation results in a "shrinkage" of the true validity when the sample equation is applied to the population. In a comparison of several estimators of R_c^2, Kennedy (1988) "demonstrated the accuracy of Stein's formula for estimating the mean of the distribution of all possible cross-validated correlations from the population from which the sample was selected." Kennedy (1988) states, Stein's Operator "could be expected to yield estimates as good as or better than cross-validation, or several other formula estimators." Stein's Operator is presented in Equation 10. In a PCSM study, Carretta & Ree (1993b) employed Stein's Operator on correlation coefficients corrected for range restriction. In

this case, Stein's Operator resulted in insignificant (no more than 0.002) reductions in the corrected correlations.

$$R^2_c = \frac{(N-1)(N-2)(N+1)(1-R^2)}{(N-p-1)(N-p-2)N} \quad , \qquad \textbf{(10)}$$

where N is the sample size and *p* is the number of predictors in the model.

2.3 Pilot Candidate Selection Process

2.3.1 AFOQT Scores

The Officer Training School (OTS) and Reserve Officer Training Corps have used the Air Force Officer Qualifying Test (AFOQT) to evaluate officer commission candidates since 1957 (Skinner & Ree, 1987). "The Air Force Officer Qualifying Test is a paper-and-pencil multiple aptitude battery used to select civilian or prior service applicants for officer precommissioning training programs and to classify commissioned officers into aircrew specialties such as pilot or navigator" (Carretta & Ree, 1995). The AFOQT is comprised of 16 individual tests that are designed to determine an applicant's abilities in five different categories: verbal, quantitative, pilot, navigator, and academic. A summary of the tests that make up the categories is provided in Table 2, followed by a description of the tests in each category as described by Carretta & Ree, 1995.

Table 3. Composition of AFOQT Composites (Carretta & Ree, 1995)

	Composite				
Test	Verbal	Quantitative	Academic Aptitude	Pilot	Navigator-Technical
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

- *Verbal Category*: Verbal Analogies measures the ability to reason and recognize relationship between words. Reading Comprehension assesses the ability to read and comprehend paragraphs. Word Knowledge provides a measure of the ability to understand written language through the use of synonyms.

- *Quantitative Category*: Arithmetic Reasoning measures the ability to understand arithmetic relationships expressed as word problems. Data Interpretation measures the ability to interpret data from graphs and charts. Math Knowledge measures the ability to use mathematical terms, formulas, and relationships.

- *Academic Aptitude Category*: The six tests that comprise the verbal and quantitative categories.

- *Pilot Category*: Verbal Analogies are described above. Mechanical Comprehension assesses mechanical knowledge and understanding of mechanical functions. Electrical Maze provides a measure of spatial ability

based on choice of a path through a maze. Scale Reading measures the ability to read scales and dials. Instrument Comprehension assesses the ability to determine aircraft attitude from illustrations of flight instruments. Block Counting measures spatial ability through analysis of three-dimensional representations of a set of blocks. Table Reading assesses the ability to extract information from tables quickly and accurately. Aviation Information measures knowledge of general aviation concepts and terminology.

- *Navigator-Technical Category*: Arithmetic Reasoning, Data Interpretation, Math Knowledge, Mechanical Comprehension, Electrical Maze, Scale Reading, Block Counting, and Table Reading are all described above. Rotated Blocks measures spatial aptitude by requiring mental rotation and manipulation of objects. General Science provides a measure of knowledge and understanding of scientific terms, concepts, principles, and instruments. Hidden Figures measures spatial ability by requiring the detection of simple figures embedded in complex drawings.

2.3.2 BAT Scores

The Basic Attributes Test (BAT) is a battery of tests administered to pilot candidates during the application process. The BAT is designed to measure a candidate's psychomotor skills, information processing, and an activity interest survey. Studies by Carretta (1989, 1990, 1992a) have validated its use in pilot selection. The test is administered with an alphanumeric keypad, monochrome computer monitor, and two joysticks at Air Force specified test facilities. The BAT is comprised of five tests including Two-Hand Coordination (psychomotor), Complex Coordination (psychomotor), Item Recognition (information processing), Time Sharing (psychomotor), and Activities Interest Inventory (attitudes) (AETC 1998). AETC's (1998) description of each test follows:

- *Two-Hand Coordination* is a pursuit tracking task (Fleishman, 1964). An airplane (target) moves in a fixed, elliptical pattern at a varying rate. The participant controls the horizontal and vertical movement of a "gun sight" using the right and left control sticks. The participant's task is to keep the gun sight on the target. The scores are summed horizontal (PS8X1) and vertical (PS8Y1) tracking distance error. These scores are then transformed to provide

a single, continuous two-hand coordination score that ranges from 0 to 25549 with lower being better.

- *Complex Coordination* measures multi-limb coordination (Fleishman, 1964). Using a dual-axis right control stick, participants are required to keep a 1-inch cross-centered on a dotted line cross that bisects the screen horizontally and vertically. Simultaneously, using the left single-axis control stick, participants have to keep a 1" vertical bar horizontally centered at the base of the screen (i.e. rudder). The scores are horizontal (PS8X2), vertical (PS8Y2), and rudder (PS8Z2) tracking error. Each of these continuous scores ranges from 0 to 72000 with lower being better.

- *Item Recognition* measures short-term memory and is based on a task proposed by Sternberg (1966). A string of 1 to 6 digits is presented on the screen. The digit string is then removed and, after a brief delay, replaced by a single digit. The participant's task is to remember the digit string and indicate whether the single digit was one of those presented in the digit string. Item recognition results in 2 variables; ITMR and ITMP. ITMR is a continuous score that ranges from 0 to 2742.8 with lower being better while ITMP is a continuous variable ranging from 0 to 100 with higher being better.

- *Time Sharing* measures the ability to perform 2 dissimilar tasks at the same time (i.e. time sharing ability; North & Gopher, 1976). In the first 10 minutes of the test, the participant is required to keep a randomly moving gun sight on an airplane (the target) using the right-handed control stick. In the next 6 minutes, the participant has to perform the tracking while simultaneously canceling digits that appear at random intervals and locations on the screen. Digit cancellation is timed and consists of pressing the same digit on the numeric keypad. The final 3 minutes of the test involve tracking only. Tracking difficulty is varied by increasing or decreasing the control stick sensitivity as a function of tracking error. The reported score is called TMSD. TMSD is a continuous value that ranges from 0 to 341.33 with higher being better.

- *Activities Interest Inventory* provides two scores, which cannot be elaborated upon due to issues related to test compromise. The two scores are AIAR and AIAP. AIAR is a continuous value that ranges from 0 to 9322.9, with lower being better, while AIAP is a discrete value that ranges from 0 to 98.77.

2.3.3 Pilot Selection Processes Across Pilot Sources

There are seven points of entry to U.S. Air Force pilot training. These include the Air Force Academy (AFA), Reserve Officer Training Corps (ROTC), Officer Training

School (OTS), Active Duty (AD), Air National Guard (ANG), AF Reserves, and international sources (Weeks, 1998). Weeks (1998) performed a comprehensive pilot selection policy capturing study for four of seven points of entry, which accounted for 65% of the pilots selected for training in that data set. The sources included in the study where AFA, ROTC, OTS, and AD. In this study, Weeks provides a detailed review of the selection process at each of the four sources. Weeks (1998) states that all selection boards perform rank lists of pilot candidates and apply a cut-off dependent on the current production requirement. The entire pilot selection process, as described by Weeks (1998), is summarized graphically in Figure 2.

For analyzing the AFA, OTS, Active Duty (non-rated officers & navigators separately) selection policies, Weeks (1998) defines the significance a variable has on the selection process by defining an average sensitivity for each variable in a resultant model. Sensitivity is defined by Weeks (1998) "as the percentage change in average board rating given a 10% change in the selection variable." For example, cumulative military performance (CMP) average had an average sensitivity of 4.78% in the AFA model. This implies that a 10% increase in CMP average resulted in a 4.78% increase in average board rating. Weeks (1998) found that PCSM score is the most significant predictor only for the Active Duty applicant selection process; however, Pugh (2003) indicates that this is no longer the case. The results found by Weeks (1998) are presented in Table 4.

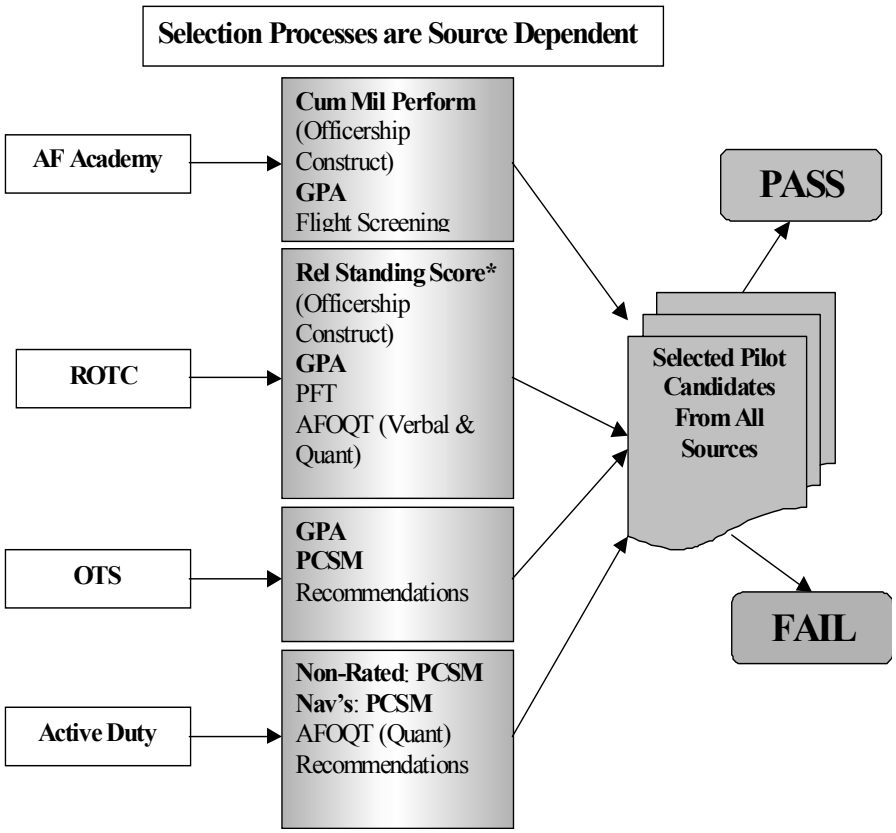


Figure 2. Pilot Selection Processes By Selection Source

For the AFA, OTS, and AD, Weeks (1998) had develop a model, which estimates these selection processes by investigating all known inputs available to the selection boards. On the other hand, the ROTC selection process is known and defined by an equation, which determines each ROTC applicant's Categorization Order of Merit (COM). Since the COM equation is known, Weeks needed only to identify which predictor variables used in calculating COM are significant. For the year studied, 1995, Weeks (1998) found 5 predictors to be significant; Relative Standing Score (RSS), GPA, PFT, AFOQT Verbal, and AFOQT Quantitative. Significance is defined by Weeks (1998) as the percentage of COM score range accounted for by each predictor. For example, RSS accounted for 47.8% of the COM score range. The other four significant

variables for the ROTC selection process and respective percentage of COM score range are presented in Table 4.

Table 4. Summary of Weeks (1998) Policy Capturing Study

AFA		OTS		AD Non-Rated		AD Navigator		ROTC*	
CMP	4.78%	GPA	1.91%	PCSM	2.09%	PCSM	1.85%	RSS	47.8%
GPA	3.21	PCSM	1.10%	2Lt Rank	-0.07%	AFOQT Quant.	0.96%	GPA	19.8%
Flight Screening Performance	1.40%	# Recommend Letters	0.43%	Positive Endorser Recommend	0.05%	Positive Endorser Recommend	0.77%	PFT	11.5%
Athletic Participation	0.07%	Interviewer Comments	0.42%	Engineer or Math Degree	0.04%	Positive Commander Recommend	0.33%	AFOQT Verbal	11.5%
Military Cmdr Position	0.04%	# Traffic Violations	-0.18%	Master's Degree	0.03%			AFOQT Quant.	9.4%
Lower Mil. Position Held	-0.03%	Possess Bachelor of Arts Degree	-0.01%	Flying Instrument Rating	0.02%			* Percentage for ROTC are % of COM Score Range	

The main results of the Weeks study are two fold. First, although the AFA and ROTC combined to provide 54% of pilots selected for training at that time of the study, measures of ability are not the most significant selection criteria for either source (Weeks, 1998). Weeks (1998) found that the AFA selection process is dominated by two factors, Cumulative Military Performance (Officership) and Cumulative Academic Average (see Table 4). ROTC selection is dominated by a measure called Relative Standing Score (RSS), which is an "Officership" score adjusted to account for differences in class size across all ROTC detachments (Weeks, 1998). Second, Weeks (1998) showed that correlation (validity) between ROTC's RSS and the pass/fail criterion was 0.01 ($p > 0.05$, $N= 469$), while the AFOQT Pilot composite had a validity of 0.14 ($p < 0.01$, $N= 469$). In

the current data, AFOQT Pilot has a correlation of 0.18 with the binary criterion. Correcting for range restriction increases the correlation to 0.21. Weeks (1998) concluded that although validity of the AFOQT Pilot composite was lower than in previous studies, it is still a much better predictor of UPT performance than measures of Officership. These results "indicates little or no relationship between Officership (*as measured by ROTC via RSS*) and pilot training attrition."

Despite Weeks (1998) findings coupled with the fact that PCSM has been shown to be a valid predictor of pilot training success in several studies (Carretta, 1992a, 1992b, 2000), the AFA and ROTC have yet to implement PCSM scores as significant factors in their respective selection processes. AETC Studies and Analysis Squadron, the sponsor of this research, promotes the increased use of PCSM in the pilot selection processes. To date, ROTC uses PCSM minimally, while the AFA has yet to implement PCSM scores (Pugh, 2003).

The AFOQT pilot composite and PCSM score have both been shown to be valid predictors of pilot training attrition (Weeks, 1998). Carretta & Ree (1992) showed that measures of ability are valid predictors of pilot training performance. Weeks (1998) found that the "AFA and ROTC pilot candidate ability levels are lower on the average than what they would be if selection policies assigned equal importance to Officership and ability." Although measures of Officership have not been shown to be valid predictors of pilot training performance, Weeks realizes the importance of leadership and responsibility in a military setting (Weeks, 1998). Therefore, Weeks (1998) suggests a balance between measures of Officership and ability in pilot selection, rather than selection based on measures of Officership that currently exists at the AFA and ROTC.

Given the results of the Weeks (1998) study, Carretta (2000) recommends implementation of a minimum qualifying PCSM score and then applying the "whole-person" concept to make final selections. Carretta (2000) noted that applying minimum qualifying PCSM scores of 25 or 50 to the 1,268 pilot trainees who successfully completed T-37 training as of Fall1998, increased T-37 graduation rates from 80.1% to 84.5% and 89.3%, respectively. If those not meeting the minimum PCSM standards are replaced by applicants who do, one could expect an increase in graduation rates. Currently, no minimum PCSM score is required for pilot training selection. There are however, minimum qualifying AFOQT standards for UPT applicants (Weeks, 1998).

For the current data set, Figure 3 presents the number of selections by PCSM score quartiles and source of selection (ROTC, OTS, AD, other). In Figure 4, failures due to training deficiency (FTD) are captured by source within each PCSM score quartile. Although AFA representation is too small in the current data to make realistic statements about the AFA selection process, one can see that the distribution of ROTC selections across the PCSM score range is different than the other sources. Specifically, the majority of passes selected by ROTC have PCSM scores less than 50. The smallest selected proportion of ROTC passes have PCSM scores greater than 75. Likewise, the failures selected by ROTC are more prevalent at the lower PCSM scores. This demonstrates the fact that PCSM is not a significant factor in the ROTC selection process.

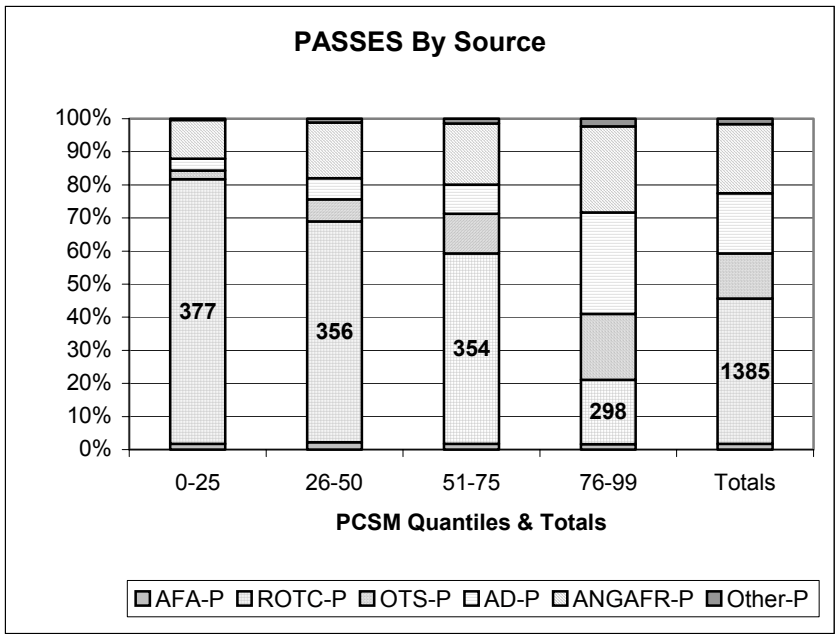


Figure 3. UPT Passes by Source

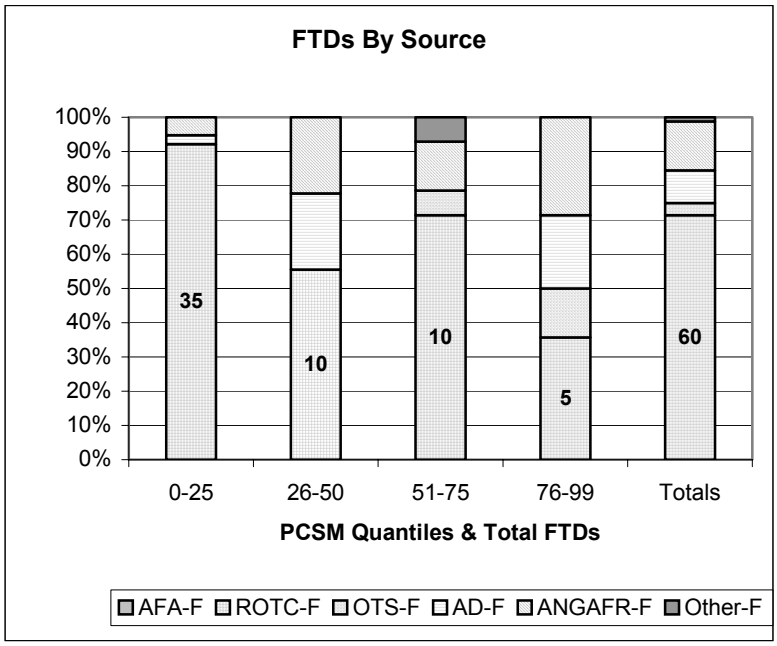


Figure 4. Failures Due to Training Deficiency by Source

2.4 Air Force and Navy Pilot Selection Model Validation Studies

2.4.1 PCSM Validation Studies

Pilot Selection Methods (Carretta & Ree, 2000a) summarizes the history of military and commercial pilot selection programs, addresses "pitfalls" in ability research, and reviews recent validation studies on AFOQT and BAT tests. Pitfalls of ability research were discussed in section 2.2. Validation studies performed for the PCSM model are discussed here. Carretta and Ree (2000a) include mission-readiness, operational tempo demands, retention issues, training costs, and safety among the reasons for sustaining a formal, scientifically based pilot selection program. As cited earlier, the cost of pilot training is quite expensive; therefore, pilot candidate attrition is of utmost importance for both military and commercial flying organizations. Reducing attrition rates through optimal selection decisions can "reduce training cost, improve job performance, and enhance organizational effectiveness" (Carretta & Ree, 2000a).

In order to ensure the maximal probability of success for each pilot applicant selected, it is necessary to make selections based on predictors known to have predictive power much greater than a naive or random type of selection. Formal validation of predictors also ensures that selection is not based on predictors that are negatively correlated with the criterion, which could introduce the possibility of test compromise by encouraging poor performance on some tests. Carretta & Ree (2000a) consider validity to be "the most fundamental testing and selection issue" (Carretta & Ree, 2000a).

Carretta & Ree (1994) performed a validity study on the predictors incorporated in PCSM. "Regression analysis was used to determine which variables provided the best prediction of two flying criteria: pass-fail flying training, and class rank at the end of

flying training" (Carretta & Ree, 1994). The AFOQT "is a good measure of general intelligence (g)" (Carretta & Ree, 1994) and has "g-saturation" of 41% (Carretta & Ree, 1995). Recall that *g* is considered to be the best predictor of training performance by Carretta & Ree.

Carretta & Ree's (1994) validation study of the variables in the current PCSM model to determine which best predict training success found that "almost all variable types were statistically significant predictors of the criteria." This suggests that there are no unnecessary variables included in the current PCSM model. However, results presented in section 4.3.4 suggest that only 3 of the 7 inputs may be significant in terms of PCSM performance. Results in Section 4.3.4 contradict this suggestion. The AFOQT Pilot composite is the single most predictive variable in the PCSM model. Carretta & Ree (1994) also investigated the incremental validity of the BAT subtests (psychomotor, information processing, risk) and flying experience above that already accounted for by the AFOQT Pilot composite. Flying experience was found to have the most incremental validity, while BAT information processing displayed the least (Carretta & Ree, 1994).

Despite the potential for using flying experience to effectively screen applicants, Carretta & Ree (1994) caution against over-weighting flying experience because of the equally likely potential for screening out successful candidate's due to "lack of income or opportunity to pursue flying training" on their own. Furthermore, although "the influence of early flying skills on later flying skills is very strong for both sexes," flying experience becomes less significant for predicting training success for later stages of UPT (Carretta & Ree, 1995, 2000a). This suggests the advantage of entering training with previous

flying experience decreases as training progresses due to learning curve effects for students who enter UPT with little or no flying experience.

Carretta & Ree (1995) revalidated AFOQT composites for validity in predicting pilot training performance using 7,563 men and women selected for pilot training, stating that "no studies have closely examined its validity for predicting pilot training performance since Miller (1966) investigated AFOQT composites." AFOQT "reflects a consensus among trainers, pilots, and researchers as to the important aptitudes for the prediction of pilot success" (Carretta & Ree, 1995).

Carretta & Ree (1995) point out that Miller did not correct correlations for the effects of range restriction. Carretta & Ree (1995) employed Lawley's (1943) correction formula to correct for range restriction. "The Lawley procedure estimates the correlations, variances, and means of both predictors and criteria as they would be found in the unrestricted population" (Carretta & Ree, 1995). They also corrected for unreliability using communalities as computed in principal factor analysis as estimates of reliabilities. Such communalities provide a lower bound estimate on reliabilities (Ree & Carretta, 2002); therefore the correction for unreliability would be based on a conservative estimate of the true reliability. The study demonstrated that "on average, the restriction in range was such that the variances in the sample were about 68% of the population variance values" for the AFOQT composites (Carretta & Ree, 1995). Carretta & Ree (1995) did not address whether Miller's (1966) study selected a sub-optimal set of predictors as a result of using uncorrected correlations, which is the point of correcting correlations during model development.

In another study, Carretta & Ree (1994) show that range restriction due to pilot training selection caused 14 of the 16 variances for AFOQT subtests to "decrease on average to 70% of the applicant sample variance" used by Skinner & Ree (1987). Each of the 16 separate AFOQT tests provided some form of predictive value in Carretta & Ree's (1995) analysis, "which leads to the conclusion that the individual composite scores are valid predictors worthy of consideration in the current analysis" (Young, 2002). In short, Carretta & Ree's (1995) study "show(s) that AFOQT is valid for the selection of pilots." Similar to Young (2002), this research considers all five AFOQT composites. Given the presence of predictive value of all 16 AFOQT sub-tests, perhaps an investigation into the creation of a new PCSM specific composite of AFOQT sub-tests would yield a replacement for the Pilot composite.

Carretta (1992b) found that "use of a training criterion based on flying performance data would not necessarily have resulted in a lower attrition rate than if the dichotomous UPT final outcome criterion was used." The data available for this research include several UPT performance measures. Although a continuous UPT performance measure is not used as a criterion in this research, the data is available as a result of the data preparation process. The data prepared for this research makes it possible for a future researcher to confirm Carretta's (1992b) results. In doing so, a composite similar to the RANKIND composite used by Carretta (1992b) could be tailored to account for the fact that no data on total flying hours completed is available in the current research data.

Carretta (1992b) studied several rank composites with UPT eliminees included in the sample. Carretta (1992b) also noted that the "criterion used in the regression had

little effect on the order ranking of the applicants once the predictors were held constant." Due to the presence of skewness that is commonly seen in AFOQT and BAT scores (i.e. high sample means with low variation) among selected pilot training candidates, Carretta (1992b) also investigated regression on log-transformed data. This resulted in "nearly identical results" (Carretta, 1992). In short, regression on a composite of actual UPT performance scores as the criterion did not change the resulting rank order of pilot candidates vs. regression on a binary pass/fail criterion.

In 1997, Carretta & Ree investigated high attrition rates among enlisted U.S. Air Force members in training for the job of weapons director. They found the failing group did not lack in ability when compared to the passing group, in fact "there were no notable differences in ability between those who successfully completed training and those who failed to complete training for non-academic reasons" (Carretta & Ree, 1997). Of 32 failures, only 3 were for academic reasons. The average score on the study's general ability composite for these 3 failures is 55, which translates to a predicted training grade of 87; a passing grade. Carretta & Ree (1997) suggest that when lack of ability cannot be identified as a cause for training failures, increasing ability standards will not reduce the attrition. In this case, they noted that all 32 failures had been non-volunteered to weapons director training. Hence, lack of motivation seemed to be the most likely cause of attrition.

It is suspected that rather than opting for self-initiated elimination (SIE) from UPT, pilot candidates sometimes purposely fail graded flying-related measures. However, the current data set includes approximately equal numbers of FTD's and SIE's, which suggests that such actions are not as pervasive as some theorize. General cognitive

ability, *g*, captured by the AFOQT pilot composite is the best predictor of pilot training success within the PCSM model; however, it is impossible to predict failures such as self-elimination because they are not ability related. No predictive model based on ability can account for a phenomenon such as lack of motivation or fear of flying. At AETC's request, the current research is conducted on a data set that includes FTD's, SIE's, and academic failures.

Weeks (1998) theorized that attrition is a function of three factors; student quality, the ratio of production to training resources available to students, and training difficulty. McLaughlin (1996) hypothesized a relationship between attrition and production. Weeks (1998) found the historical average attrition rate to decrease at times of low pilot production and increase in average attrition rate at times of high pilot production. When production quotas increase, resources are generally not increased at all or at a proportionate level (Weeks, 1998). This causes pilot training resources to become scarce, thus driving up the production to resources ratio.

Although Weeks (1998) hypothesizes that training complexity is increasing, due to the complexity of modern cockpits and introduction of mission oriented training, a relationship between attrition and training complexity has yet to be shown. Weeks (1998) states that the current de-emphasis of ability at AFA and ROTC may be combining with an increase in production to resource ratio, to increase attrition beyond what would be seen if more emphasis in selection was put on ability. Further, Weeks (1998) predicts further increases in attrition if selection policy continues to focus on non-ability measures.

The fact that Weeks (1998) found that attrition rate is associated with USAF pilot production quotas may imply that the standard by which the UPT pass/fail criterion is based is not constant. Of course, other factors include possible changes in selection processes and quality of those selected in years of high production quotas. Thus, the data includes records of UPT failures during years of higher pilot production, which may have been a pass in years of low production. The opposite case is also likely, where passes in years of high production may not have been selected over failures selected in years of low production. Furthermore, changes in production quotas cause changes in the distributions of predictive test scores for selected pilot candidates. Surely, such artifacts in the data make discrimination more difficult, thus limiting the predictive power of the resultant models.

2.4.2 Validation of Naval Aviation Tests

Williams et al. (1999) performed a revalidation of the Aviation Selection Test Battery's (ASTB) utility for predicting performance in naval aviation ground school and flight training grades. The ASTB is the Navy equivalent of the AFOQT. The ASTB was originally introduced in 1942 and the current version dates back to 1992 (Williams et al., 1999). Despite not correcting sample correlations for range restriction, the results of the ASTB validation were positive. Williams et al. (1999) provide the following description of the ASTB. The ASTB consists of six paper-and-pencil sub-tests, which are used to generate three composites used in the naval pilot selection process. All three have been validated to predict their intended criterion. An academic qualification rating (AQR) predicts ground school performance. A Pilot Flight Aptitude Rating (PFAR) predicts flight grades in primary flight training. The Pilot Biographical Inventory (PBI) predicts

attrition through primary flight training. Although Williams et al. (1999) found the ASTB to perform well at predicting ground school (academic) and flight training grades; it provided the ability to predict attrition "to a lesser extent". Hence, it appears that attrition is not only related to flying and academic abilities.

Furthermore, Williams et al. found that a relatively new computer-based performance test (CBPT), similar in description to the BAT, shows promise as a tool for selection of U.S. naval aviators. The CBPT has yet to be implemented in the naval pilot selection process (Williams et al., 1999). Although, only a small sample (N=210) of data was available, CBPT performed well as a predictor ($R^2 = 0.33$, $p < 0.0001$). The sample (200 male, 10 female) volunteered to take the CBPT prior to beginning Aviation Pre-flight Indoctrination (API). The CBPT data provided incremental validity beyond that of using PFAR alone, by accounting for 17% of the primary flight grade variance in the sample (Williams et al., 1999). See Williams et al. (1999) for a more detailed description of the CBPT and its 10 sub-tests.

Although only 15% of those who take the ASTB are selected for naval aviator training, Williams et al. (1999) specifically chose not to correct correlations for range restriction in the revalidation study. Despite this, Williams et al. found correlations of moderate magnitude. These correlations were comparable to historical ASTB findings of Frank & Baisden (1993) and Hiatt et al (1997), which are cited by Williams et al. (1999).

Damos (1996) also refrained from correcting for range restriction. Damos found results similar to Williams et al. (1999) for a wide variety of aviation selection tests. The merits of Damos' argument against correcting for range restriction are considered in section 2.2.1.6. Williams et al. (1999) did not consider CBPT tests requiring "rudder

pedals or more than one joystick" citing reliability, calibration, and quality control problems with more complicated psychomotor test batteries. As a side note, the USAF BAT currently includes tests requiring more than one joystick and a new version of the BAT requiring the use of rudder pedals is nearing operational use (Pugh, 2003). Despite the presence of possible unreliability in the BAT scores related to the EQPMOT input to PCSM, results of this research presented in chapter 4 suggest that EQPMOT or its component scores has been shown to be the most significant scores among the BAT sub-tests.

2.5 Current PCSM Database and T-37 Performance Data

The research sponsor, AETC/SAS, provides the data used in this research. Predictive & demographic data is contained in a PCSM database. UPT performance data is contained in separate SSN and MASS databases. The SSN database provides data on the actual UPT pass/fail outcome, while the MASS database provides UPT performance scores. The data preparation process used to generate the consolidated data set used in this research is described in Section 3.3.

The current data set contains 3,343 records, 3,155 are passes and 188 failures. The data includes 3,086 males and 257 females. Table 5 presents the breakout of passes and failures by selection source in the current data set. Table 6 presents UPT outcome by sex. Table 7 presents the number of each type of failure contained in the data set by sex. Figure 5 present the information of Table 7 as percentages of each failure type within each sex.

Table 5. Pass/Fail Breakout by Source

Source	AFA	ROTC	OTS	AD	ANG/AFR	Other	Totals
Passes	56	1385	428	575	658	53	3155
Failures	0	134	8	24	21	1	188
Total	56	1519	436	598	779	55	3343

Table 6. Breakout of UPT Outcome by Sex

	Pass	Fail	Total
Female	225	32	257
Male	2930	156	3086
Total	3155	188	3343

Table 7. Number and Type of UPT Failures by Sex

Failure Type	FTD	Academic	SIE	Total
Female	15	2	15	32
Male	70	12	74	156
Number of Failures	85	14	89	188

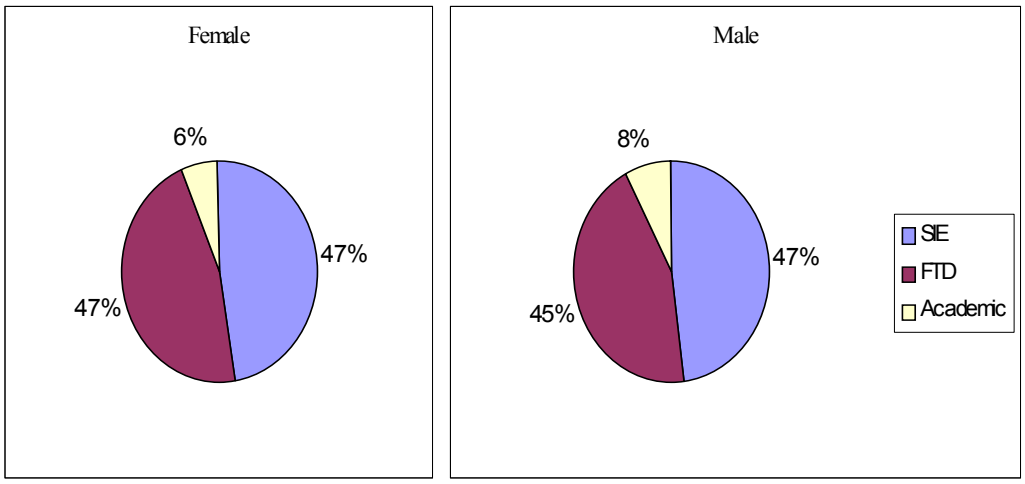


Figure 5. Proportion of Failure Types by Sex

2.6 Factor Analysis

Dillon & Goldstein (1984) describe factor analysis as a "data reduction technique for investigating interdependences." Dillon & Goldstein (1984) differentiates factor analysis methods from other data reduction methods such as principal component analysis by stating that factor analysis techniques "distinguish different types of variance," rather than simply accounting for total variance. In general, unobservable factors are associated with a set of observable variables, which represent a common or shared type of variation. Factor analysis attempts to find links between seemingly unrelated variables to a common factor structure. For the common factor-analytic model, "interest centers on that part of the total variance that is shared by the variables" (Dillon & Goldstein, 1984). The variables "linked" together for each factor are then used to make interpretations about the latent structure underlying the data.

In exploratory factor analysis, the researcher seeks to investigate, interpret, and ultimately understand the factors underlying the data. Linkages between observable

variables and unobservable factors are established via factor loadings. Under certain circumstances, factor loadings represent the correlation between the variables and the factors (Dillon & Goldstein, 1984). Confirmatory factor analysis, on the other hand, seeks to confirm or deny the hypothesized interpretation of the underlying factor structure. In this case, the dominant factors are thought to be understood and represented by the variables used in the PCSM model. For example, PCSM's 7 inputs originate from 3 primary sources; the AFOQT, the BAT, and flying experience. This research uses factor analysis to confirm the linkages between these inputs and the most dominant underlying factors.

2.7 Discriminant Analysis

Discriminant Analysis attempts to discriminate between two or more groups within a population. This is done by deriving a discriminant function that when applied to independent predictors, classifies each exemplar as a member of one of the groups. The discriminant function is a linear combination of independent variables. Discrimination is accomplished by maximizing between-group variance relative to the within-group variance" (Dillon & Goldstein, 1984). Once applied to an individual exemplar, the discriminant function assigns a score on the discriminant function line. This discriminant score is "essentially a weighted average of the exemplar's values on a set of independent variables" (Dillon & Goldstein, 1984). After all exemplars are assigned a discriminant score, an a posteriori probability of the likelihood of belonging to each group can be derived for each score. Deriving two distributions along the discriminant function line does this. Figure 6 is a graphical representation of a two-group discriminant analysis from Dillon & Goldstein (1984).

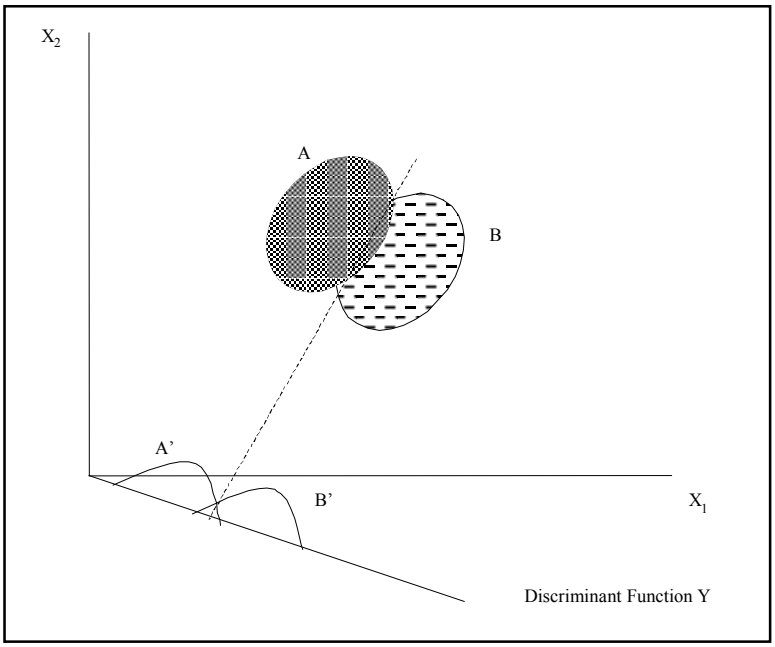


Figure 6. Graphical Illustration of Two-Group Discriminant Analysis

2.7.1 Discriminant Analysis Methodology

Discriminant Analysis is typically based on the assumption the data is multivariate normal and that the variance-covariance matrix of independent variables in each group is the same (Young 2002). Fisher's original derivation of the linear discriminant function did not specify any distributional assumptions (Dillon & Goldstein, 1984). Dillon & Goldstein (1984) provide a summary of work done to test the robustness of the linear discriminant function to departures from multivariate normality and equality of variance-covariance matrices. Optimal results are obtained when these two assumptions hold; however, many studies have ignored the assumptions based on the fact that Fisher made no distributional assumptions (Dillon & Goldstein, 1984).

Dillon & Goldstein (1984) caution that the linear discriminant function is not robust enough to ignore the two assumptions and give the following summary of their

findings in the literature. They found that studies on the effects of unequal variance-covariance structures indicate that the test for the equality of group mean vectors is adversely affected. When multivariate normality is violated, "tests of significance and estimated classification error rates may be biased." Normality should be investigated if estimated error rates are greatly different for the groups in the population. To test equality of covariance matrix for the two-group problem, the Box M method is used. Details concerning the development and implementation of the Box M method are available in SPSS's online manual (SPSS 2002a). A significant p-value for the Box M method implies that the variance-covariance matrices obtained for the two groups are not equal, thus rejecting the null hypothesis. SPSS provides the option to use separate variance-covariance matrices in its discriminant analysis procedure. Some caution the using the Box M test of equal covariance structure stating, "the multivariate Box M test is particularly sensitive to deviations from multivariate normality, and should not be taken too 'seriously'. (StatSoft Inc., 2003).

If equality of variance-covariance structure holds, then Fisher's approach is applicable. Fisher showed that Equation 11 gives the vector of discriminant weights ($\hat{b}$), here S_p is the pooled sample covariance matrix shown in Equation 12 (Dillon & Goldstein, 1984). Here $\bar{x}_i$ is the centroid of group x_i and n_i is the number of exemplars in the i^{th} group.

$$\hat{b} = S_p^{-1} \cdot (\overline{x_1} - \overline{x_2}) \qquad \textbf{(11)}$$

$$S_p = \frac{1}{n_1 + n_2 - 2} \cdot (x_1^T x_1 + x^T{}_2 x_2) \qquad \textbf{(12)}$$

The discriminant score for each individual is obtained via the linear combination of the value measured for the independent variables and the associated discriminant weights. Discriminant scores are obtained via $Y = \hat{b}^T X$, where X is a *p x n* matrix of *p* predictors and *n* exemplars (Dillon & Goldstein, 1984).

In the case of unequal variance-covariance structure, Dillon & Goldstein (1984) provide a discussion of the performance of a quadratic discriminant function developed by Smith (1947). The reader is directed to this discussion for details concerning the use of quadratic discriminant function and it's performance relative to Fisher's linear discriminant function under the same conditions.

One can test whether the between-group differences in average score profiles of the two groups are statistically significant. This is can be accomplished via an F-test of the test statistic Z in Equation 13, where D^2 is the Mahalanobis generalized distance (Dillon & Goldstein, 1984). Dillon & Goldstein (1984) state that Z has an F-distribution with *p* and $n_1 + n_2 - p - 1$ degrees of freedom if the hypothesis of equal means and a common variance-covariance matrix holds.

$$Z = \frac{n_1 n_2}{n_1 + n_2} \cdot \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2) \cdot p} \cdot D^2 \qquad \textbf{(13)}$$

Classification of exemplars is done via a classification rule, which defines a cut-off score on the discriminant function, **Y**. If group sizes are equal, Y is generally defined as the midpoint between the average discriminant scores for each group. If group sizes are unequal, a weighted cut-off score, $\mathbf{Y}^*$, will optimize classification error within the data set used to derive the discriminant function. However, when the group representation is significantly disproportionate, this can result in all exemplars being classified as a member of the group represented by the larger group. This defeats the purpose of classification, especially when the target group is the smaller group. Equation 14 gives the expression for a weighted cut-off score provided by Dillon & Goldstein (1984). A graphic representation presented by Dillon & Goldstein (1984) and taken from Young (2002) in Figure 7 presents the placement of Y and Y^*.

$$Y^* = \frac{n_2\overline{Y}_1 + n_1\overline{Y}_2}{n_1 + n_2} \qquad \textbf{(14)}$$

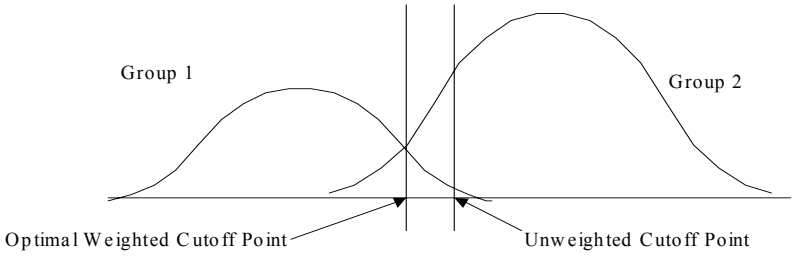


Figure 7. Optimal cutting score with unequal sample sizes

Once all exemplars have been classified, it is desirable to estimate the classification error rate. Common methods include apparent error rate (APER), holdout method, and cross-validation method. If all exemplars are used in deriving the discriminant weights, APER "estimates are consistent but can be severely optimistically biased" (Dillon & Goldstein, 1984). APER is an estimate of the combined rate of misclassification for both groups. A *confusion matrix* makes calculating APER straightforward. Table 8 presents a *confusion matrix* and Equation 15 presents the APER equation.

Table 8. Confusion Matrix

		Predicted Membership	
		Group 1	Group 2
Actual Membership	Group 1	N_{1C}	$N_{2\bar{C}}$
	Group 2	$N_{1\bar{C}}$	N_{2C}

$$APER = \frac{N_{1\bar{c}} + N_{2\bar{c}}}{N_{1C} + N_{1\bar{c}} + N_{2C} + N_{2\bar{c}}} \quad \textbf{(15)}$$

The holdout method allows for validation by simply splitting the exemplars into two randomly chosen groups of predetermined sizes (i.e., 2/3 and 1/3). The larger group is used to determine the discriminant weights, which are then applied to the smaller holdout group. This gives a better estimate of how the discriminant function will perform in the population because the holdout group has not been introduced to the model during

development of the discriminant weights. In this research, the previously discussed "TEST" set is the holdout group.

Cross-validation is accomplished by holding out the i^{th} exemplar and determining a discriminant function based on the *N-1* remaining exemplars. The held out i^{th} exemplar is then classified as in the holdout procedure. This process is repeated for all $i = 1 \ldots N$ exemplars. APER is then calculated based on the resulting classification of the single holdout exemplar across the *N* iterations of the cross-validations process. Most statistical applications provide this type of cross-validation as an option.

2.7.2 Stepwise Discriminant Analysis

Stepwise Discriminant analysis is commonly used when there are many predictors available for use in determining group classification (Dillon & Goldstein, 1984). The use of partial F-values and probability of F-value are common methods determining the most important predictors for discriminating between groups. A partial F-value is conditioned on only those predictors present in the discriminant function at the present step, and not the entire set of $p - 1$ predictors as in the nominal sense of an F-value (Dillon & Goldstein, 1984).

The stepwise discriminant analysis process is similar to stepwise multiple regression analysis as presented by Dillon & Goldstein (1984) in chapter 6 of that text. The following process reflects Dillon & Goldstein's (1984) summary of how to conduct stepwise discriminant analysis.

1. First, single predictor F-values are computed, treating each variable as though it were the only predictor available.

2. The predictor with the largest F-value is then chosen to enter the discriminant function.

3. Successive steps add (or delete) new predictors on the basis of their computed F-values conditioned on those predictors already made part of the system.

Dillon & Goldstein (1984) also note that recent work suggests "liberal α-levels" for F-to-enter values. It is suggested to use $0.10 < \alpha < 0.25$, rather than conventional levels of $\alpha < 0.10$. Furthermore, Dillon & Goldstein (1984) states that stepwise discriminant analysis suffers from the same problems discussed for multiple regression analysis in chapter 6 of their text. The two main issues reported by Dillon & Goldstein are the affects of multicollinearity that results from including strongly correlated predictors and the fact that partial F-values are such that the F-distribution does not strictly apply. Refer to pages 240-242 of the Dillon & Goldstein (1984) text for a more complete discussion.

Dillon & Goldstein (1984) also present a method of canonical discriminant analysis. In some cases, this method is preferred over Fisher's linear discriminant functions. The coefficients for both Fisher's linear discriminant functions and canonical discriminant function coefficients are available for each discriminant model developed in this research from AETC/SAS.

2.7.3 Arguments Against Stepwise Methodology

Whitaker (1997) provides a detailed review of some of the written commentary of several researchers who "sharply criticized" the use of stepwise methodologies and provides alternative suggestions. Several researchers are cited for their support of

multiple linear regression as superior to discriminant analysis in most situations. Specifically, Whitaker (1997) cites Kerlinger (1986) and Thompson's (1994) criticism of transforming a continuous criterion into a dichotomous criterion in order to use discriminant analysis because of the valuable variance information that is "squandered." Whitaker cites Thompson (1989, p. 166) as stating the "it has not been shown that package stepwise results are relevant for a predictive discriminant analysis," where group classification is the point of the analysis. This is pertinent as group classification is exactly what is intended in the current research.

Whitaker (1997) cites multiple researchers who have "challenged traditional interpretations of statistical significance." It is argued that popular statistical packages use incorrect degrees of freedom in statistical tests built into computer programs that do discriminant analyses (Whitaker, 1997). Secondly, sampling error can represent the only differences in predictors. Thus, stepwise procedures can erroneously capitalize on these differences in sampling error. Likewise, "otherwise worthy variables are often excluded from the analysis altogether and assumed to have no explanatory or predictive potential" (Whitaker, 1997). Finally, due to the previous two problems, it is argued that the stepwise methodology often fails to select the best subset of variables. Whitaker suggests correcting computer generated F statistics by hand and conducting the "all-possible-subsets" approach to determining the best sub-set of variables to overcome the above mentioned problems.

2.8 Logistic Regression

Logistic regression is used as on of the means of updating the regression weights of the current PCSM model. Currently, PCSM is based on a linear regression that is transformed to in a logistic sense by way of applying a discrete sigmoid function approximation to the linear regression output. Logistic discrimination can be used in situations where measurements have been collected on both quantitative and qualitative predictors. If multivariate normality and common variance-covariance structure holds, posterior probabilities of membership in the i^{th} group conditioned on the current exemplar can be expressed as follows in the *multivariate logistic function* (Dillon & Goldstein, 1984):

$$\begin{aligned} P(G_1 \mid x) &= \exp(\beta_0 + \beta^T x) \cdot P(G_i \mid x) \\ P(G_2 \mid x) &= \frac{1}{1 + \exp(\beta_0 + \beta^T x)} \end{aligned} \qquad \textbf{(16)}$$

where β are the logistic regression coefficients. A logistic model is distinct from a linear regression model in that the outcome is dichotomous. When the logistic distribution is used, a common notation for the conditional mean of the criterion given a predictor score is given in Equation 17.

$$\Pi(x) = E(Y \mid x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \qquad \textbf{(17)}$$

The *logit transformation* of $\Pi(x)$ is central to the study of logistic regression because it "has many of the desirable properties of a linear regression model" (Hosmer & Lemeshow, 1989). Specifically, the *logit*, $g(x)$ is linear in its parameters. Another critical difference between linear and logistic regression is that the error term in logistic regression will take on one of two possibilities due to the dichotomous nature of the logistic output. The two possibilities are presented in Equation 18.

$$\begin{aligned} y = 1 &\Rightarrow \varepsilon = 1 - \pi(x) \quad \text{w.p.} \quad \pi(x). \\ y = 0 &\Rightarrow \varepsilon = -\pi(x) \quad \text{w.p.} \quad 1 - \pi(x) \end{aligned} \tag{18}$$

Hosmer & Lemeshow (1989) and Bauer (2002b), as well as most multivariate texts with sections dealing with logistic regression, provide detailed development of the fitted logistic regression model. The development is based on the method of maximum likelihood to yield estimated values for unknown parameters, which maximize the probability of obtaining the observed set of data (Hosmer & Lemeshow, 1989). The method of maximum likelihood involves construction of the *likelihood function*, which expresses the probability of the observed data as a function of the unknown parameters. It is common practice to take the natural log of the *likelihood function* as a first step, known as the *log likelihood*. The log likelihood is then differentiated with respect to the parameters β_0 and β_1 (for the two parameter model). The derivatives are then set equal to zero and solved for the respective parameters. Solving the equations for the logistic regression requires iterative methods because the equations are non-linear in the parameters. The equations to be solved, which can be found in most logistic regression

texts, are presented in Equation 19 and most statistical applications provide the capability to find their solutions.

$$\frac{\partial}{\partial \beta_0}(L) = \sum_{i=1}^{N} \left[y_i - \pi(x_i) \right] = 0$$
$$\frac{\partial}{\partial \beta_1}(L) = \sum_{i=1}^{N} x_i \left[y_i - \pi(x_i) \right] = 0 \qquad \textbf{(19)}$$

The logit of the multiple logistic regression model and the resulting likelihood equations follow the same form as in the two-parameter model, with the addition of multiple beta coefficients to be estimated.

2.8.1 Interpretation of the Coefficients of the Logistic Regression Model

Log of the odds ratio is called the log-odds ratio or just log-odds, which is the logit difference. It approximates a quantity called relative risk. The odds ratio will tend to have a skewed sampling distribution "due to the fact that it is bounded away from zero" (Hosmer & Lemeshow, 1989). Hence, inferences are usually based on a sampling distribution of ln(odds ratio) = $\overline{\beta}_1$, which tends to follow a normal distribution.

It is recommended to code all dichotomous variables as 0-1 and treat them as interval scaled, because other coding schemes affect the estimate of the odds ratio and the end points of associated confidence intervals (Hosmer & Lemeshow, 1989). The most common interval scaled coding method uses a referent group (vs. the method of deviation from the means coding used in linear regression) because of the interest in estimating the

risk of an "exposed" group relative to a control group or unexposed group (Hosmer & Lemeshow, 1989).

2.8.2 Two Parameter Logistic Regression Model for Personnel Selection

A study by Raju et al. (1991) of 84,808 observations of Air Force enlistees tested on forms 8,9,and 19 of the Armed Services Vocational Aptitude Battery (ASVAB) showed the logistic regression model to be "valid and also quite robust with respect to direct and indirect range restriction on the predictor" for a two parameter logistic model. A dichotomous criterion was created using information on Final School Grade (FSG) for the population. Since data was only available for those passing their respective Air Force school (FSG ≥ 70), the criterion for receiving a label of "success" was set at FSG ≥ 84, the median FSG grade in the data population. The two predictors used were Math Knowledge and Mechanical Comprehension scores. The two parameter model used by Raju et al. (1991) is presented in Equation 20, where "D is a constant usually set to 1.7 to make P(x) correspond to a normal ogive and *a* and *b* are job parameters to be estimated." Here, D does not denote Mahalanobis distance.

$$P(x) = \frac{\exp[Da(x-b)]}{1+\exp[Da(x-b)]} \quad \textbf{(20)}$$

One benefit of the two-parameter logistic regression model is that the results can "directly relate the probability of job success to trait levels" (Raju et al., 1991). Advantages of the logistic model cited by Raju et al. (1991) include the following:

1. The standard error of an observed correlation coefficient does not vary from one predictor score to the next, whereas the standard error of P(x) depends on x. Therefore, the information that logistic regression provides about the precision of measurements in more useful.

2. Because logistic regression is used in item response theory, P(x) can be considered to be subpopulation invariant, whereas range restriction is known to affect the correlation coefficient.

Raju (1991) randomly selection of 1,000 samples of 1,000 observations each from the population (N = 84,808) and performed logistic regression. The results show that the two-parameter logistic regression model's theoretical probabilities for the entire dataset fit the empirical probabilities reasonably well. Performance with respect to direct and indirect range restriction was also studied. Direct range restriction was induced via a cut-off set at the population median Math Knowledge score for each population sample. Indirect range restriction was induced for set of samples by setting a cut-off score at the population median Mechanical Comprehension prior to applying the cut-off for Math Knowledge. Here the sample consisted of those whose Mechanical Comprehension and Math Knowledge scores were both greater than the respective population medians. Under indirect range restriction, 3% of the χ^2 values were significant at the same alpha level. Range restriction reduced average sample size from n = 1,000 to 527 for direct restriction and 565 for indirect range restriction.

A χ^2 test was employed to assess how well the sample-based logistic regression probabilities matched the sample-based empirical probabilities. At the $\alpha = 0.01$ level, 2% of the 1,000 χ^2 values were significant for unrestricted samples. Likewise, 2% of the samples subjected to direct range restriction were significant at the same alpha level, while 3% were significant for indirect range restriction. This shows that the effects of

range restriction did not significantly effect the resulting logistic regression model goodness of fit to the population based logistic regression model. The author agrees with Raju et al.'s (1991) statement, "overall, the two-parameter logistic regression model appeared to offer a promising alternative to studying the question of the probability of success in selection." However, updating the regression weights for the current PCSM model involves 6 parameters including the intercept.

2.9 Artificial Neural Networks

Artificial Neural Networks (ANN) is inspired by the architecture and function the human brain. Figure 8 presents a "node" of the McCullock-Pitts model (1943). This was the first attempt at modeling such brain functions (Looney, 1997). The terms node and neurode are used in the literature interchangeably. The node is the building block of a neural network. Nodes simulate the biological neuron, which functions via synaptic inputs that activate an output depending on whether or not the sum of the inputs to a single neuron exceeds some threshold of the neuron (Looney, 1997). In a neural network, a user specified number of nodes receive an input from each predictor variable for an exemplar presented to the network. In the biological sense it is believed that the brain learns which synaptic inputs to a specific neuron should be given more weight in determining its resultant output. Activation functions transform a linear combination of weighted predictor inputs to form the node's output.

The set of optimal weights applied to the predictor inputs must be learned. The McCullok-Pitts model does not include a mechanism for the model to *learn*; however,

this basic model served as the impetus for future networks (Looney, 1997). Methods for simulating learning in a neural network model are discussed in Sections 2.9.4 and 2.9.5.

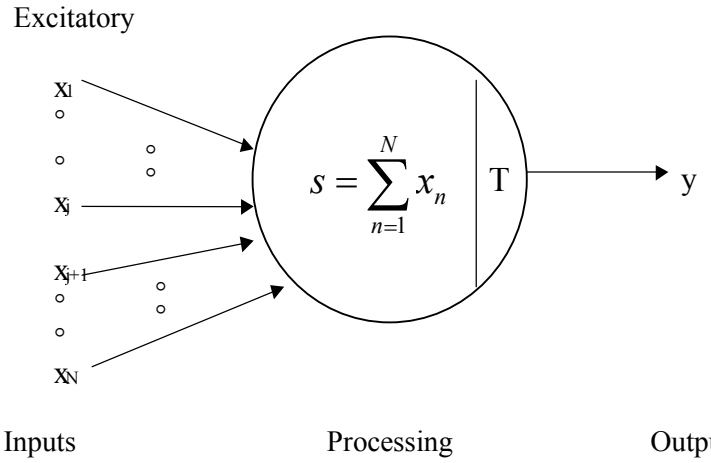


Figure 8. McCulloch-Pitts neuronal model

Bi-valued threshold functions, unipolar (0,1) and bipolar (-1,1), where the first activation functions. These are step-functions; however, continuously differentiable activation functions are most common in current applications. The advantage these functions offer is that gradient methods can be used to iteratively solve for weights that map a vector of input features (predictors) into a desired output that matches its class identifier (Looney, 1997). The sigmoid or logistic function is a unipolar (0,1) example of a continuous activation function commonly used today. The equation for this function and its graph are presented in Equation 21 (Looney, 1997) and Figure 9 (Young, 2002), respectively. The hyperbolic tangent function (bipolar sigmoid) is its bipolar (-1,1) continuous counterpart, whose equation and graph are presented in Equation 23 (Looney, 1997) and Figure 10 (Young, 2002), respectively.

$$f(s) = \frac{1}{1 + \exp(-\alpha(s - b))}, \qquad \textbf{(21)}$$

where s is the weighted sum of input features, α is the decay (growth) rate, b is the bias that shifts the function center to where e^0 occurs (at s = b), where the output is the midvalue f(s) = ½. Hence, b is the s-axis center of asymmetry of f(s) (Looney, 1997).

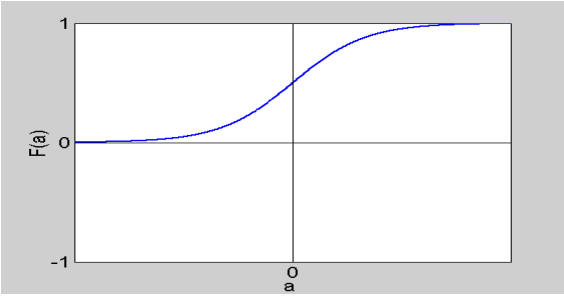


Figure 9. Sigmoid Activation Function

$$f(s) = \frac{1 - \exp(-\alpha s)}{1 + \exp(-\alpha s)} = \tanh\left(\frac{\alpha s}{2}\right), \qquad \textbf{(22)}$$

where the last term is derived in the usual situation by setting the threshold equal to zero (Looney, 1997).

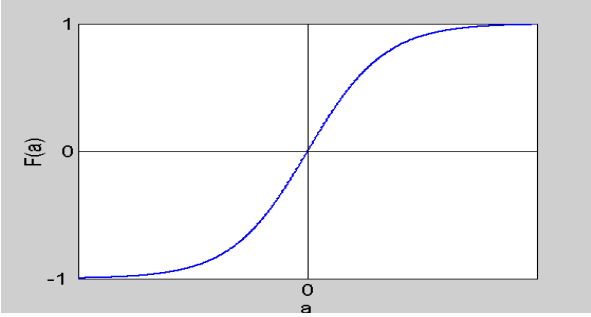


Figure 10. Hyperbolic Tangent Activation Function

2.9.1 Artificial Neural Network Definitions

The following definitions are related to artificial networks and are often referred to in discussions of neural networks and are taken from Bauer (2002). The definitions are included as an aid to the reader.

- Activation Function. A mathematical function that maps the sum of the weighted values entering a node into a range of output values (Looney 1997).
- Artificial Neural Network (ANN). An information processing system (algorithm) that operates on inputs to extract information and produces outputs corresponding to the extracted information (Bauer 2002b).
- Architecture. The topological arrangement of neurons, layers, and connections, which defines the set of modeling equations available to the ANN (Bauer 2002b).
- Backpropagation. A learning algorithm for updating weights in a feed-forward multi-layer perceptron (MLP) ANN that minimizes the mean squared mapping error (Bauer 2002b).
- Conjugate Gradient Method. A weight updating method that measures the gradient of the error surface after each pass. It then alters the weights of the node inputs using a compromise between the direction of the steepest gradient and the previous direction of change (SPSS 1997).
- Epoch. A complete presentation of the data set being used to train the MLP, or equivalently called a training cycle (Bauer 2002b).

- Feature. In neural networks, features refer to the input vectors of information, which are presumed to have some relation that may be helpful in distinguishing the various output classes. The vector of features is often called an *exemplar* (Bauer 2002b).

- Feedforward Neural Network. Multilayer ANNs whose connections exclusively feed inputs from lower to higher levels. In contrast to a feedback or recurrent ANN, a feed-forward ANN operates only until all the inputs propagate to the output layer. An example of a feed-forward ANN is the MLP (Bauer 2002b).

- Hidden Units. The processing elements in a MLP ANN that are not included in the input or output layers. This is part of the neural network located between the input and output layers (Bauer 2002b).

- Hyperbolic Tangent Activation Function (Tanh). An activation function that maps a node's inputs to a continuous S-shaped function between –1 and 1. The continuous function allows the network to utilize gradient search methods for the weight updates (SPSS 1997). Figure 10.

- Learning Algorithm. The equations used to modify the weights of processing elements in response to input and output values (Bauer 2002b).

- Linear Activation Function. An activation function that simply sums the inputs to a node and passes them through (SPSS 1997). Figure 9.

- Neuron. The fundamental building block of an ANN. Normally, each neuron takes a weighted sum of its inputs to determine its net input. The net input is then processed through its transfer function to produce a single-valued output that is broadcast to 'downstream' neurons (Bauer 2002b).

- Sigmoid Activation Function. An activation function that maps a node's inputs to a continuous S-shaped function between 0 and 1. The continuous function allows the network to utilize gradient search methods for the weight updates (SPSS 1997). Figure 10.

- Supervised Training. A method of training adaptive ANNs that requires a labeled training set and an external teacher. The teacher knows what the desired response is and thus can provide responses for correct or incorrect classification by the network (Bauer 2002b).

- Weight. The values associated with each connection in the network that signifies the importance of the respective inputs. The weights are combined to calculate the activations (Bauer 2002b).

2.9.2 Development of the multilayer perceptron (MLP) model

Neural networks "learn" by adjusting the weights that are applied to the input features, which ultimately affect the output(s) of the network. Development of the Rosenblatt perceptron in the 1950's was significant because it included multiple neurodes in a two-layer network (Young, 2002). A drawback of the Rosenblatt perceptron is that it could only effectively solve linearly separable classification problems (Looney, 1997). The XOR problem, a classic non-linearly separable classification problem involving two groups, required another neural network innovation before it could be solved effectively. The solution to such a problem was an algorithm commonly known as backpropagation. Backpropagation, developed by Rumelhart, Hinton, & Williams in 1986, "utilizes gradient search of the error space to update the weights of the network" (Young, 2002). This allowed for a relaxation of the exponential number of perceptrons required to solve non-linear problems, which hampered the development of neural networks prior to the development of computing power capable of handling the computational costs associated with the number of nodes needed (Looney, 1997).

A two-layer network has an uncounted input layer of features, a hidden layer of neurodes that accept a weighted sum of all inputs into each neurode, and an output layer that accepts weighted sums from each of the hidden nodes (Looney, 1997). Figure 11 presents a generalized picture of the common feedforward ANN with two layers. The hidden layer is made up of unipolar or bipolar sigmoidal threshold logic functions (TLF), which provide binary responses as inputs to the output layer. The output layer is also driven by TLF's. Each output node produces a binary response, which then can be interpreted as a means of classifying exemplars.

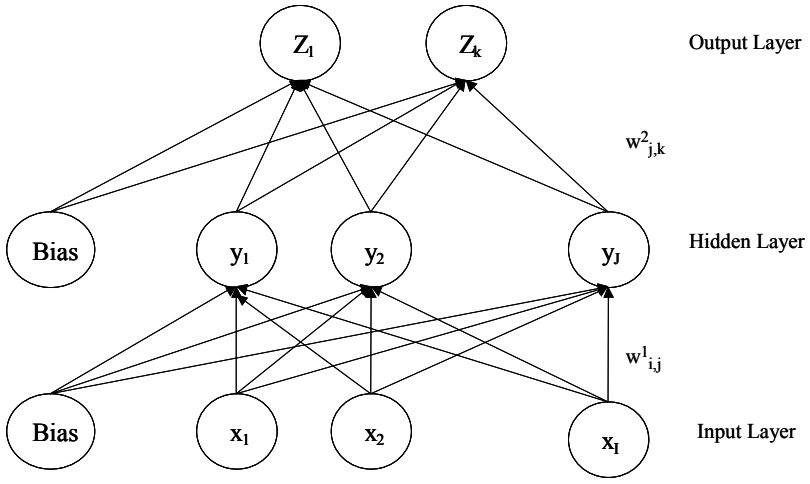


Figure 11. General Feedforward MLP

2.9.3 Network Engineering

Much of the architecture of current networks rely on the Hornik-Stinchcombe-White theorem, which states (Looney, 1997):

> A feedforward artificial neural network with two layers of neurodes and nonconstant nondecreasing activation function at each hidden neurode can approximate any piecewise continuous function from a closed bounded subset of Euclidean N-dimensional space to Euclidean J-dimensional space with any prescribed accuracy, provided that sufficiently many neurodes be used in the single hidden layer.

Hence, although it is possible to solve non-linearly separable classification problems with a single layer network, the previous theorem states that the two-layer MLP is sufficient. The process involves the use of a sufficient number of neurodes in the hidden layer to separate the classes into linearly separable sub-groups and then using the network output layer to adjoin or lump the sub-groups into the appropriate classes. The following is a

review of the main ideas presented by Looney (1997), who provides a more detailed discussion of network engineering.

A set of non-linearly separable classes can always be decomposed into linearly separable subclasses because each neurode defines a hyper-plane, which separates the feature space into two half-spaces. The intersection of these hyper-planes creates convex hulls, which are pair wise linearly separable (Looney, 1997). The feature space can be divided into a sufficient number of convex hulls in order to classify the data set into linearly separable subgroups; even if it means classifying the exemplars as individual singletons. As the dimensionality of the data increases, the required number of neurodes increases exponentially (Looney, 1997). Too many neurodes causes over fitting or specialization of the network to the data set, which causes poor generalization upon cross-validation or application to a new sample (Looney, 1997). Specialization occurs when a large number of weights allow the network to essentially "memorize" the data set. Generalization is enhanced when successive layers have less neurodes than the one previous (Looney, 1997).

Looney (1997) states that smaller networks have the following advantages; (1) better generalization, (2) learn more quickly, and (3) operate with less complexity requiring less computer memory. Looney cites Villiers & Barnard (1992) for showing that two hidden layers are always sufficient. The second hidden layer is often referred to as the output layer. By implementing a second hidden layer, the number of neurodes required in the first hidden layer is reduced, thus improving the networks generalizability (Looney, 1997).

A common approach to network design involves the addition of one or more noise features. Bauer et al. (2000) implement a random uniform (0,1) noise feature as a means of identifying salient features in their signal-to-noise ratio (SNR) method. This method is discussed in section 2.9.6. Holmstrom and Koistinen (1992) added noise features to transform an underdetermined system into an over determined one (Looney, 1997). Young (2002) found that initializing network weights from a Gaussian distribution (± 0.001) outperformed networks initialized with weights from a uniform distribution (± 0.001). Both uniform and Gaussian distributed initial weights were investigated in this research and Young's (2002) findings were not supported. Therefore, the more commonly used uniform weights were used to initiate networks.

2.9.4 Backpropagation

This research will utilizes a conjugate gradient method of network training; however, backpropagation is reviewed because it is the most common training method. Backpropagation allows a feedforward MLP to iteratively update the network weights, thus training the network to classify the target output vector. Gradient search methodology, which involves iteratively calculating derivatives of the network's error surface with respect to the network inputs and respective hidden layer weights, systematically reduces the error between the network's output and the desired output. There are two methods of updating network weights. Batch training updates the network weights after the entire set of exemplars in the data set are passed through the network. The total error calculated is then used to perform the error surface derivates.

Instantaneous training updates the network weights after each individual exemplar passes through the network. Bauer (2002) prefers instantaneous training.

In chapter 11 of his text, Looney (1997) provides a robust discussion of many feedforward architectures and weight updating methods. Topics covered include gradient descent vs. strategic search algorithms, fullpropagation vs. backpropagation, and the effects and comparisons of a multitude of MLP algorithms. Due to over correction of the network weights with the steepest descent method, the backpropagation algorithm for this research incorporates the conjugate gradient method for updating the network weights during instantaneous training. The conjugate gradient method is discussed in Section 2.9.5.

2.9.5 Conjugate Gradient Method

The conjugate gradient method uses a compromise between the direction of steepest descent and the previous direction of change (SPSS 1997). This helps the network avoid the problems of overcorrecting weights encountered by Young (2002). The conjugate gradient method takes advantage of the fact that close to the "well" of a local or global minimum, the Total Sum-Squared Error (TSSE) function of network error is approximately quadratic, so that convergence can be completed with a fixed number of steps, which eliminates a major problem of backpropagation (Looney, 1997). Although, the problem of finding such minimums in the feature space still remains, strategic search methods can be used to locate a starting weight set in the region of a "deep minimum" prior to implementing the conjugate gradient method (Looney, 1997).

Johansson et al. (1992) state "backpropagation is likely the most common supervised learning algorithm in neural network applications." Unfortunately, backpropagation suffers from becoming inefficient for training network weights. As the number of weights in the network increases, backpropagation learning time can become prohibitive. The conjugate gradient method is a numerical optimization technique that has been shown to reduce learning rates in backpropagation type weight training (Johansson et al., 1992).

The usual methods of overcoming the problem of excessive learning time exhibited by backpropagation include reducing the dimensionality of the problem, using faster computers or a parallel processing architecture, applying numerical optimization techniques. The conjugate gradient method falls in the latter category. Johansson et al (1992) list the following advantages of the conjugate gradient method.

1. Faster than backpropagation's steepest descent method by an order of magnitude on the parity problem
2. Doesn't suffer from inefficiencies and possible instabilities caused by using a fixed step size in steepest descent
3. Simple for a numerical optimization technique compared to second order Newton and Quasi-Newton methods due to avoidance of calculation and inversion of the Hessian matrix of second order partial derivatives of the error surface at each iteration

Calculation of the Hessian matrix is a common problem in numerical optimization. Complexity of Hessian calculation and inversion grows exponentially as problem dimensionality increases. Storage requirements for large Hessian matrices are also cause for concern, even with today's computing power. Fortunately, the conjugate

gradient method avoids calculation of the Hessian. Johansson et al. (1992) develop the conjugate gradient method in detailed fashion.

In order to compare the two methods, Johansson et al. (1992) implemented both methods to 4 and 5 bit parity problems across a series of learning rate/momentum combinations for each of 6 different network architectures (3 with a single hidden layer, 3 with two hidden layers). The result was that for each problem, "the conjugate gradient methods are an order of magnitude faster than conventional backpropagation." However, Johansson et al. (1992) noted "the relative successes of optimization algorithms are highly problem dependent." Networks with a single hidden layer resulted in more convergence failures than two hidden layers for both the conjugate gradient and conventional backpropagation methods.

2.9.6 Signal to Noise Ratio

Insignificant or non-salient input features adversely affect the classification accuracy of an ANN. *Signal to Noise Ratio* (SNR) is a method proposed by Bauer, Alsing, & Greene (2000) for screening out non-salient features in multi-layer perceptron feedforward ANN's. This method is an improvement over earlier methods used by Belue & Bauer (1995) and Steppe & Bauer (1996) because the current method of feature screening can be done with only one iterative training run. The previous methods required between 10-30 training runs to iteratively remove non-salient features. Belue & Bauer (1995) used a "partial derivative based saliency measure to calculate each feature's effect on a single hidden layer ANN's output." Steppe & Bauer (1996) used Tarr's

weight-based saliency measure by "summing the squared values of the weights connecting feature *i* to the hidden nodes."

The SNR saliency measure in Equation 23 involves converting a ratio of first layer weights of a fully trained ANN to a decibel scale (Bauer, Alsing, & Greene, 2000). The ratio in Equation 23 sums the squared first layer weights from the candidate input features of a trained ANN and the first layer weights from an injected noise input that follows a Uniform (0,1) distribution.

$$SNR_i = 10\log_{base10}\frac{\sum_{j=1}^{J}\left(w^1_{i,j}\right)^2}{\sum_{j=1}^{J}\left(w^1_{N,j}\right)^2}, \qquad \textbf{(23)}$$

where

- SNR_i is the saliency metric for the i^{th} feature
- J is the number of hidden nodes
- $w^1_{N,j}$ is the weight connecting the injected noise feature, x_N, to the first hidden node layer
- $w^1_{i,j}$ is the weight connecting the input feature, x_i, to the hidden node layer

The study by Bauer, Alsing, & Greene (2000) showed that the SNR feature screening method is "consistent and robust within and across most ANN architectures." In the study, architectures with momentum rates of 0.1, 0.5, and 0.9 were investigated.

Inconsistency was noted for architectures with high momentum rate set at 0.9. SNR_i, the SNR saliency measure of feature *i*, is used to rank order the features by significance to the network with larger saliency measures being more salient. An SNR of 0.0 indicates the feature has the same saliency as the noise feature, which is not related to the target in any way. The least-salient feature is removed at the end of each training epoch during a single training run until all are removed. Classification error is record for each training epoch. The first salient feature whose removal causes a significant increase in classification error and all features removed after that are then retained and used to fully train a new network. A modified version of this SNR method developed by Young (2002) is presented in Chapter 3 and used in this research

2.10 Ensemble Method

Perrone & Cooper (1992) presented an *ensemble method* for combining multiple neural networks, possibly of "different architectures or trained on different data sets", to improve performance of the combined model above any of the individual models. The presence of many local minima that generally exist in the weights of neural network makes simple averaging of the weights, as is done in the parameter space for most resampling techniques, counter productive (Perrone & Cooper, 1992). The ensemble method averages in "functional space not parameter space," which allows it to actually benefit from the presence of local minima captured by the different networks created (Perrone & Cooper, 1992). One drawback of the ensemble method; however, is requirement to develop and maintain multiple dissimilar models so that their outputs can be combined.

Two methods are presented by Perrone & Cooper (1992), the Basic Ensemble Method (BEM) and Generalized Ensemble Method (GEM). Figure 12 is reproduced from Perrone & Cooper's work to provide an illustration of how the ensemble method improves the solutions in a Mean Square Error (MSE) sense. In figure 12, Regions A and B represent two distinct classes. Hyper-planes 1 and 2 represent two possible models. Hyper-plane 3 is an improved solution, in the MSE sense, resulting form averaging hyper-planes 1 and 2 and "will give the optimal generalization performance" (Perrone & Cooper, 1992). Perrone and Cooper also provide an example that illustrates the weakness of the BEM, which the interested reader should consider.

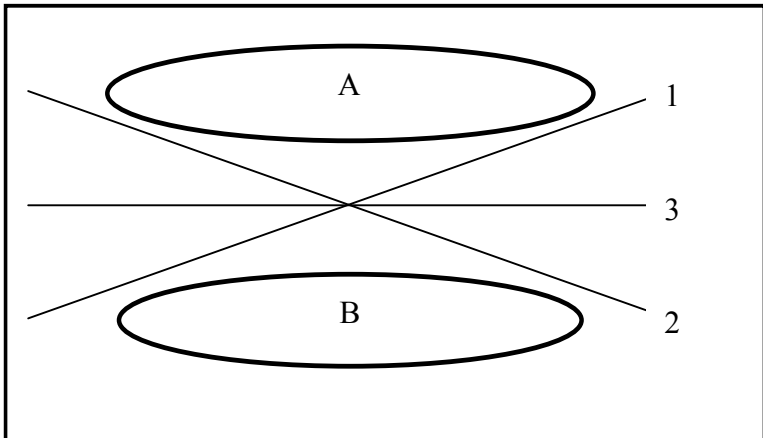


Figure 12. Ensemble Illustration

Perrone & Cooper (1992) contend that the usual method of creating multiple networks and selecting the best performing network based on some criteria discards valuable information contained in the other networks. Another important benefit of the ensemble method is its ability to use all available data in training each network, which

avoids the problem of missing data common to cross-validation. The ensemble method's "smoothing property of the ensemble process removes any over-fitting" (Perrone & Cooper, 1992).

In addition to neural network problems, the ensemble method is applicable to any technique that minimizes MSE (Perrone & Cooper, 1992). Benefits of the ensemble method described by Perrone & Cooper (1992) include:

1. Efficient use of all networks of a population without discarding any network

2. Efficient use of all available data without over fitting

3. Inherent performance of regularization by smoothing in functional space which helps avoid over-fitting

4. Utilizes local minima to construct improved estimates, rather than being hindered by local minima

The GEM "always generates a regression estimate which is ... the best possible of any linear combination of the elements of the population of functions which estimate *f(x)*" (Perrone & Cooper, 1992). The Generalized Ensemble Method combines N networks such that:

$$f_{GEM}(x) \equiv \sum_{i=1}^{i=N} \alpha_i f_i(x) = f(x) + \sum_{i=1}^{i=N} \alpha_i m_i(x), \qquad \textbf{(24)}$$

where

- $f(x)$ = the true but unknown network function
- $f_i(x)$ = the trained network function
- $\alpha_i's$ are real numbers that satisfy the constraint $\sum \alpha_i = 1$.
- $\alpha_i = \frac{\sum_j C_{ij}^{-1}}{\sum_k \sum_j C_{kj}^{-1}}$, which minimizes MSE[$f_{GEM}$] = $\sum_{i,j} \alpha_i \alpha_j C_{i,j}$
- $C_{ij} = E[m_i(x)m_j(x)]$
- $m_i = f_i(x) - f(x)$

C_{ij} is the correlation matrix of the output errors from the different networks. The above results from Perrone & Cooper rely on two assumptions: Linear independence of the rows and columns of C_{ij} and a reliable estimate of the true correlation matrix C_{ij}. By forming the correlation matrix between the different networks, one is able to calculate "weights" to apply to the output of each net. Summing the weighted outputs of each network, forms a new model that reduces the MSE of the overall model (Young, 2002). If the models combined via the GEM method are not dissimilar enough the weights derived become uniform; hence the GEM becomes similar to the BEM. Perrone & Cooper caution that, in practice, if two or more networks are not dissimilar, the correlation matrix C will be ill-conditioned. Perrone & Cooper also noted that in their sample the magnitude of the increase in classification accuracy was reduced after 6-8 networks where included in the GEM.

Young et al. (2003) performed network screening by forming a matrix of the errors from the *n* different networks generated. An *n x n* correlation matrix of these

errors was calculated and factor analysis was performed to identify which networks to retain. To actually identify the networks, Young et al. (2003) identified the factor structure by employing the common practice of retaining all factors with eigenvalues greater then 1.0. The network with the highest factor loading for the respective principal factors used in the factor analysis where retained for use in the Generalized Ensemble Method (Young et al., 2003).

2.11 Chapter Summary

This chapter discussed the following broad areas; PCSM research, pitfalls of ability research, the pilot candidate selection process, validation studies of the current PCSM model, an overview of the data used is in this research, factor analysis, and several predictive techniques that are employed in this research. The techniques described are discriminant analysis, logistic regression, artificial neural networks, and the ensemble method of combining predictive models.

In terms of the "pitfalls," the only pitfall that is not understood in terms of possible impact on the results of this research is range restriction. Correlations corrected for range restriction are calculated as part of the validation study. Despite the fact that many studies have investigated the accuracy of the correction under, no study reviewed addressed a specific example where an invalid model resulted from using uncorrected correlations. Carretta & Ree (1995) revisited Miller's (1966) study, but did not challenge the selection of the variables in Miller's (1966) model.

The other pitfalls have been accounted for in this research or have been addressed through previous study results that are thought to be relevant to this research. For

example, corrections for unreliability proved to be of trivial significance (Carretta & Ree, 1994). Dichotomization of the criteria had little effect when replaced with a more continuously scaled criterion (Carretta, 1992b). Sample sizes in this research are more than adequate for the indirect range restriction situation studied by Sackett & Wade (1983). Subgroup effects are not expected to affect factor analytic results (Carretta, 1997 and Carretta & Ree, 2000a). Stein's operator yielded little change in correlations corrected for range restriction (Carretta & Ree, 1994). Furthermore, the issue of cross-validation is addressed by the use of the independent TEST set.

Review of the PCSM model reveals that UPT candidates are selected based on a variety of selection policies dominated by different selection criteria. PCSM is not a significant factor in the selection of a majority of UPT candidates. This forces PCSM validity to be based on its performance as a predictor for candidates whose selection is not necessarily based on their PCSM score. Nevertheless, PCSM has still been shown to be a valid predictor of successful completion of UPT.

III. Methodology

3.1 Introduction

This chapter provides an overview of the data available for this research, the process of preparing the data for analysis, the methodology and the algorithms employed to accomplish the analysis, and the method for comparing competing models. Furthermore, a brief description of two specialized software applications is presented. The methodology described in this chapter is used to answers the three objectives of in this research. First objective is PCSM validation. Secondly, the regression weights of the current PCSM model are updated. Finally, an independent model is developed.

The two specialized software applications used in this research are SPSS 11.5 and SPSS Neural Connections 2.1. SPSS 11.5 is a statistical analysis package capable of performing many popular multivariate analysis techniques. Neural Connections 2.1 primarily supports the development of neural network models. The descriptions of these packages provide a short tutorial for readers who wish to recreate parts of this research. Model performance and comparison is accomplished primarily through the presentation of Receiver Operating Curves (ROC). A ROC is useful for classification problems because performance is displayed across the entire range of decision thresholds, thus allowing the users to select the threshold that best meets their needs.

3.2 Data Description

The data for this research was provided by AETC/SAS on one CD-ROM. The data exists in Microsoft ACCESS database format. Figure 13 presents how the data was

merged into the single data table used for this research. The process and issues encountered are described in more detail in Section 3.3.

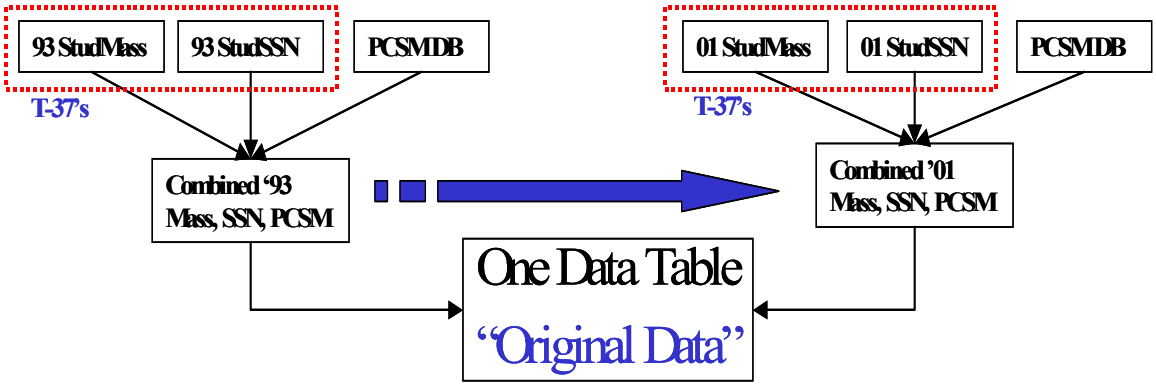


Figure 13. Data Preparation Process

The data is made up be two sets of databases. The first contains UPT performance data from 1993 through 2001. There are several UPT performance databases for each year, each of which includes multiple data tables. Based on the advice of AETC, the "hist_yr" database was selected for each year. Within the "hist_yr" database, two data tables were selected; "SSN" and "MASS." The SSN table contains data relevant to UPT success or failure. The MASS table contains UPT performance data. The second set of databases available contained PCSM data. PCSM data covers the period 1993 through 2001 in single data table, whereas the SSN and MASS data are contained in separate data tables in separate databases for each year. With the data sources selected, the UPT data contained in the SSN and MASS tables for each year were matched up with the predictive data available in the PCSM database.

3.3 Data Preparation

The analysis focuses on the data of individuals whose UPT training course is identified as one of the following: PV4AA, PV4AB, PV4AJ, and PV4AN. These courses distinguish the training as non-joint military training. All other courses are disregarded for purposes of this research. In order to merge the data into one baseline data set, the three data sources were matched across the SSN, MASS, and PCSM data within each year. To do so in Microsoft Access, a common identifier is needed for definition of query relations. Social Security Number is present in the PCSM and SSN data for all years. However, the Social Security Number field is missing from the MASS data for the years 1993-1998. The training base location and student ID fields were used in the SSN and MASS tables to append a Social Security Number field to the MASS data. To do this, a query matched student ID in the SSN and MASS tables such that the training base location field in each table matched. The second logical operator is necessary because the student ID field is only unique within training base locations.

Prior to performing queries to merge matching data fields, it was necessary to search the SSN tables for duplicate Social Security Numbers. Duplicates occur most often for people who attended one of the four courses of interest in this research as well as another type of flight training course. In all cases, the record associated with the course of interest was kept, while other records were deleted. Where duplicate records existed for a course that is not of interest, all records were deleted from the SSN table. After the 1993-1998 MASS tables had a Social Security Number field appended, a reference field that identifies year was added using an update query.

A separate data merge query was performed for each year, 1993-2001. Each query defined two relations based on the Social Security Number field. These relations matched the SSN table to the MASS table and the PCSM table to the MASS table within a single query. Criteria defined in the query limited selected records to those records where Social Security Number in the MASS table matched in the SSN and PCSM tables. Each year's query result was then exported as a data table into a separate database. A complete, stand alone data set was constructed by combining each year's query result into one data table in the new database. Only one match exists for 1993 and is missing data in many fields. It was deleted. No matches existed for 1994. A complete data set covering 1995-2001 generated 3,409 records prior to further data review.

Table 9 presents all data field labels chosen for this research. Some demographic data such as SSN, name, gender, and year are for reference only and is not used as part of any analysis. Text Nominal variables such as Status_Source, Aero_Rating, and Fly_Exp are converted to 0,1 dummy variables for each level of the text nominal variable. IFT score is not used as a predictor as it is not available to the selection boards.

Table 9. Baseline Data Variables

Data Type	Name	Scale	Description
Criterion	Pass_T37	0,1	0=Fail 1=Pass
Demographic	SSN	Nominal	Social Security Number
	Last_Name	Text	Last Name
	First_Name	Text	First Name
	Gender	0,1	0=Male, 1=Female
	DOB	Nominal	Date
	Ed_Level	Ordinal (0-6)	0=High School 1=1 Year College 2=2 Years College 3=3 Years College 4=Bachelors 5=Masters 6=PhD
	GPA	Scale (0-4.0)	Grade Performance Average
	Status_Source	Text Nominal	AFA ROTC OTS_AD OTS_Civ AD AFR ANR Other
	Year	Nominal	1993-2001
Flying Experience	Aero_Rating	Text Nominal	None Student Pilot License Private Pilot License Commercial Pilot License Transport Airline Certified
	Fly_Exp	0,1 Dummy variable for each category	0=Not Applicable 1=Applicable None Fixed Wing Rotary Wing Single Engine Multiple Engines Instructor Instrument Rating

Data Type	Name	Scale	Description
	Flt_Hr_Cd	Ordinal (0-9)	0=None 1=1-5 Hours 2=5-15 Hours 3=16-25 4=26-40 Hours 5=41-60 Hours 6=61-100 Hours 7=101-150 Hours 8=151-200 Hours 9=200+ Hours
	IFT_Score	Scale (0-99)	Initial Flight Training Score
AFOQT Data	Pilot	Scale (1-99)	Pilot Composite Score
	Nav	Scale (1-99)	Navigator Composite Score
	Quant	Scale (1-99)	Quantitative Composite Score
	Verbal	Scale (1-99)	Verval Composite Score
	Acad	Scale (1-99)	Academic Composite Score
BAT Test	H2CX1	Scale (0-72,000)	BAT Pursuit Tracking Error
	PS2X2	Scale (0-72,000)	BAT Pursuit Tracking Error
	PS2Y2	Scale (0-72,000)	BAT Pursuit Tracking Error
	PS2Z2	Scale (0-72,000)	BAT Pursuit Tracking Error
	ITMR	Scale (0-2,742.8)	Item Recognition Score #1
	ITMP	Scale (0-100)	Item Recognition Score #2
	TMSD	Scale (0-341.33)	
	AIAP	Scale (0-98.77)	Activity Interest Score #1
	AIAR	Scale (0-9,322.9)	Activity Interest Score #2
	BAT_Score	Scale (1-99)	BAT Score
	BAT_Age	Nominal	Age at time of BAT
PCSM	Raw_PCSM	Scale (0-1.24)	Regression Weighted Linear Combination of PCSM Inputs
	PCSM	Scale (1-99)	Final PCSM Score after applying a discretized sigmoidal transformation
UPT Data	Course	Text Nominal	5 Character Flight Training Course Identifier
	Last_Stat_Date	Nominal	Most Recent Status Date
	Last_Stat_Code	Text Nominal	Most Recent Status Code to identify Pass/Fail type
	Last_Stat_Phase	Nominal (1-3)	1=Academic 2=T37 Phase 3=T38 Phase
	T37_Raw_CK	Scale (0-99)	Raw Check Flight Grade
	T37_Raw_DLY	Scale (0-99)	Raw Daily Flight Grade
	T37_Raw_EPQ	Scale (0-99)	Raw
	T37_Raw_ACAD	Scale (0-99)	Raw Academic Grade

Data Type	Name	Scale	Description
	T37_Raw_FCR	Scale (0-99)	Raw Final Commander's Ranking
	T37_T_CK	Scale (0-99)	Transformed Check Flight Grade
	T37_T_DLY	Scale (0-99)	Transformed Daily Flight Grade
	T37_T_EPQ	Scale (0-99)	Transformed
	T37_T_ACAD	Scale (0-99)	Transformed Academic Grade
	T37_T_FCR	Scale (0-99)	Transformed Final Commander's Ranking
	* No Transformed UPT performance data available for 1993 and 1994		

The complete data set was scrubbed for missing or corrupted data. Missing GPA's were replaced with the mean GPA of 3.11 in 543 cases. One record was deleted for missing data in many fields. Six records were deleted for zero or empty scores for one or more AFOQT scores. Four records were deleted for zero BAT subtest scores. These 11 deletions brought the total number of records down to 3,398. The AFOQT Academic field had 42 empty records. Instead of deleting these records, the field was replaced via simple linear regression (y = ax + b). To do this, an independent variable was defined as the average of the AFOQT Verbal, Quantitative, and Navigator fields. The existing AFOQT Academic score was the dependent variable. The resulting equation for replacing missing Academic scores is presented in Equation 25 and had an R-squared of 0.946.

$$ACAD = 1.1957\left[Avg\left(V,Q,N\right)\right] - 15.225 \quad \textbf{(25)}$$

In order to generate a standard method of including an age variable, the date of birth and BAT test date fields were combined to create an age at BAT test field via Equation 26.

$$BAT\_Age = \frac{(Age@BAT - DOB)}{365.25} \quad \textbf{(26)}$$

The Last_Stat_Code and Last_Stat_Phase fields were used in an update query to create a binary Pass/Fail criterion. Those with non-elimination codes and a Last_Stat_Phase of at least 2 where defined as a pass (1.0). All others were defined as a fail (0.0). All failures with elimination codes other than E51 (Flying Deficiency), E52 (Academic Deficiency), and E56 (Self-Initiated Elimination) were deleted. This brought the total number of records down to 3,343. This included 3,155 passes and 188 failures. The retained failures included 85 for flying deficiency, 14 for academic deficiency, and 89 for self-initiated elimination.

Nominal data such as Status_Source and Fly_Exp were replaced with binary dummy variables at each of their respective levels, where "1" indicates the presence of the associated identifier. In keeping with standard convention, the number of dummy variables is one less than the total number of categorical levels. All zeros indicate the AFA for the Status_Source field. All zeros indicate rotary wing for the Fly_Exp field. Ordinal data such as Ed_Level, Aero_Rating, and Flt_Hr_Cd were coded with increasing integer values beginning with zero. Scaled data was not transformed in anyway at this point; however, standardization of all inputs is used for independent model development. No standardization of data was performed for the PCSM validation and regression update because the current PCSM model does not perform such transformations.

One specific difficulty that exists in the data as provided by AETC is the existence of inconsistencies between the Aero_Rating, Fly_Exp, and Flt_Hr_Cd fields.

For example a person identified as a commercial pilot in the Aero_Rating field may have zero reported FAA flying hours in the Flt_Hr_Cd field or indicate only single engine flying experience in the Fly_Exp field. Where the Flt_Hr_Cd is greater than zero, the Fly_Exp field is updated to indicate both fixed wing and single engine flying experience if it was not already indicated. Where reported FAA flying hours indicate between 1-40 flying hours, the Aero_Rating is updated to student pilot if the field was empty or zero. Likewise, where flying hours indicate greater than 40 flying hours, the Aero_Rating field is updated to private pilot if the field was empty or zero. Finally, instances where zero flying hours were reported and the Aero_Rating indicates student pilot or private pilot still existed in the data. In cases where flying hours indicate zero, Aero_Rating was updated to none. In cases where flying hours indicate 1-20 and Aero_Rating indicates private pilot, the Aero_Rating was updated to student pilot. AETC was consulted prior to these updates. All updates were done using Microsoft Access update queries.

This preparation process results in a data set referred to as DATA_A. It has approximately 94 % passes (3,155) and 6% failures (188). It was decided to validate the PCSM model on two other data sets with equal proportions of passes and failures. Two additional data sets with equal pass/fail proportions were created from DATA_A. DATA_B is created using bootstrap resampling. Resampling is done with replacement from the 188 original failures until a total of 3,343 failures records including the original 188 are obtained, hence DATA_B includes 6,686 records. 188 passes were randomly selected and included with the 188 original failures to create the second of these two data sets, DATA_C. These three data sets are used to validate the current PCSM model.

For the regression update and development of the independent model, the data sets change as a result of holding out the independent "TEST" data set mentioned several times thus far and increasing the failure proportion in DATA_A. However, the names given to the data sets are the same. This is because the methods for creating the different data sets are unchanged. Only the size and pass/fail proportions are changed. Hence, all references to DATA_B imply that a large bootstrap resampling has occurred to generate enough additional failure records to give equal pass/fail proportions.

The attrition rate has remained near 10% for the past several years (Pugh, 2003). AETC requested the data be made to represent a 10% attrition rate. By contrast, Young (2002) performed a similar analysis for AETC using only flying training deficiencies (FTD). In his research, FTD's alone accounted for 5% of the data set, whereas FTD's, SIE's and academic deficiencies combined account for only 4.6% of the current data set. The original PCSM model was developed using only FTD's, which represented 20% of the data set. Later, a PCSM update included all types of failures, which represented 19.6% of that data set (Carretta, 2003). How the actual numbers and types of failures in these historical data sets compare is unknown.

Again three data sets were generated for analysis in similar fashion described for the validation study. The only difference is that the TEST set is first pulled out of DATA_A before creating DATA_B and DATA_C. To create the TEST set, 25% of the 188 failures (N=47) are first removed. Then 423 passes are randomly selected so the TEST set results in the required 10% attrition rate. As a result, DATA_A is left with 2,732 passes and 141 failures. In order to force the 10% attrition requirement on DATA_A, bootstrap resampling with replacement was performed 163 times on the 141

remaining original failures and added to DATA_A. For the purposes of the regression update and developing the independent model, DATA_A now contain 2,732 passes and 304 failures. DATA_B and DATA_C sets were then generated in the same fashion described for the validation study from this new version of DATA_A. DATA_B has 2,732 passes and 2,732 failures, while DATA_C has 141 passes and 141 failures. Table 10 summarizes the sizes and pass/fail proportions in the data sets used in each phase of this research.

Table 10. Data Set Summary

	VALIDATION		LOGREG		INDEP-MODEL	
	PASS	FAIL	PASS	FAIL	PASS	FAIL
DATA_A	3,155	188	2,732	304	2,732	304
	96%	4%	90%	10%	90%	10%
DATA_B	3,155	3,155	NA	NA	2,732	2,732
	50%	50%	NA	NA	50%	50%
DATA_C	188	188	912*	304	141	141
	50%	50%	75%	25%	50%	50%
TEST	NA	NA	423	47	423	47
	NA	NA	90%	10%	90%	10%

*** 3 of these data sets were created**

3.4 Specialized Software Utilized

3.4.1 SPSS

SPSS for Windows, version 11.5, was used to perform several analysis techniques. These include basic descriptive statistics, factor analysis, stepwise linear regression, and stepwise discriminant analysis. Figure 14 shows the initial window for SPSS Linear Regression. It is similar to the setup window for most SPSS analysis functions. The general analysis setup in SPSS opens such a window for variable selection, which

includes buttons to open successive windows for setting analysis parameters. Dependent and independent variables are selected in a straightforward manner. Other parameters are selected by means of check boxes in the successive windows.

Dillon and Goldstein (1984) state, "conventional levels (< 0.10) can often tend to terminate the stepwise process prematurely." In keeping with this conservative modeling strategy, probability of F-to-enter was set at 0.10 and probability of F-to-remove was set at 0.12. Since Data_B is so large, it is necessary to set the probability of F values to the traditional 0.05 and 0.10, respectively, to limit the number of variables in the final model.

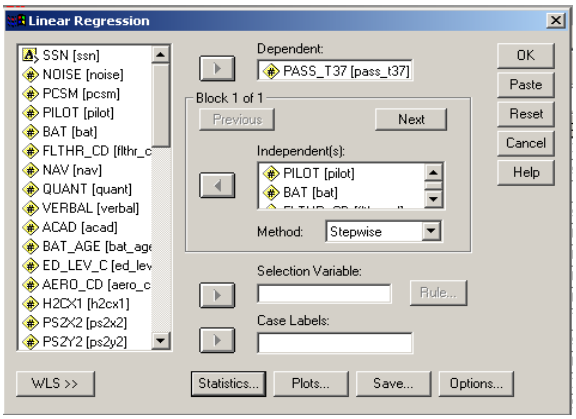


Figure 14. SPSS Linear Regression

For discriminant analysis, the method of testing predictive value in the model chosen was the Mahalanobis distance with probability of F-to-enter set at 0.10 as seen in Figure 15. The initial set up window for discriminant analysis is quite similar to Figure 14. Figure 16 displays classification options selected for the discriminant analysis. The prior probabilities option simply defines a single threshold for predicting group membership. This research utilizes ROC curves, which allow analysis of model

performance across all decision thresholds. ROC curves are discussed in section 3.9. The separate group-covariance matrices option is selected because the Box M test of the assumption of equal covariance matrices is rejected with a p-value 0.000, rounded to 3 decimal places. Doing so improves performance of the discriminant function when the equal covariance assumption is violated.

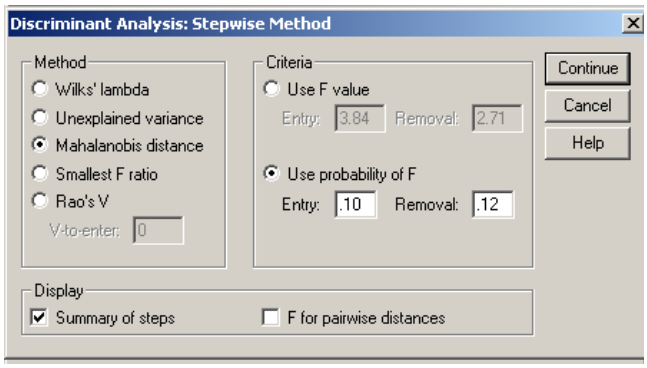


Figure 15. Discriminant Analysis Method Options

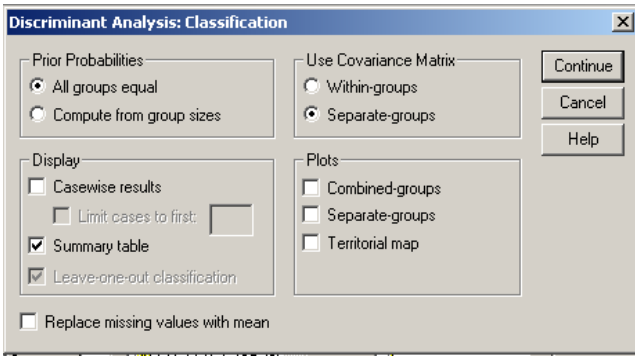


Figure 16. Discriminant Analysis Classification Options

SPSS provides a summary table in the form of a confusion matrix. For the pass/fail classification problem, a confusion matrix is simply a 2 x 2 table that records the correct and incorrect group predictions. A sample confusion matrix is presented in Table 11. Table 12 provides definitions for the confusion matrix, where True Positive indicates

correct prediction of the target group (failure) and False Positive indicates incorrect prediction of the target group. SPSS also allows the user to apply the discriminant function to a validation set. In this case, a confusion matrix for the validation set would also be displayed. Furthermore, SPSS provides the capability to perform a Cross-Validation, where an iterative process determines discriminant functions with a single exemplar held-out. Group prediction is then done via the discriminant function determined using all available data except the exemplar for which the current prediction is being made. Group prediction, discriminant scores, and group membership probabilities can be saved as additional columns in the data sheet used in the analysis.

Table 11. SPSS Confusion Matrix

Classification Results[a]

			Predicted Group Membership		
		PASS_T37	.00	1.00	Total
Original	Count	.00	34	16	50
		1.00	207	579	786
	%	.00	68.0	32.0	100.0
		1.00	26.3	73.7	100.0

a. 73.3% of original grouped cases correctly classified.

Table 12. Confusion Matrix Definitions

	Predicted Fail	**Predicted Pass**
True Fail	True Positive	False Negative
True Pass	False Positive	True Negative

3.4.2 Neural Connections

Neural Connections, version 2.1 is used in the development of all neural networks. Neural Connections is a Windows-based program, which allows the user to perform multiple types of neural networks in a relatively straightforward manner. Setting up a neural network in Neural Connections involves three main steps; importing and configuring the data, selecting the network type and its associated parameters, and setting up the output. Neural Connections opens with a blank template. The user then builds the network in a building block fashion by selecting network tools from a tool bar provided as part of the Neural Connections environment. Figure 17 presents a simple MLP network architecture, which contains the three network tools necessary to create a model.

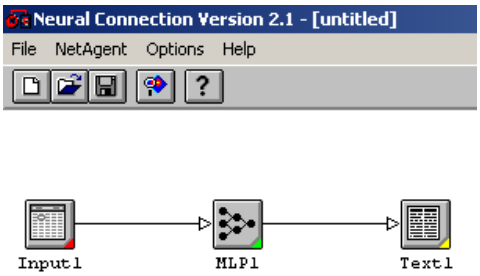


Figure 17. Sample Neural Connections MLP Architecture

After the user selects the tools for building a model, opening dialog boxes is accomplished by clicking on the tools themselves. Clicking on View within the Input1 tool menu opens an empty data table similar to a spreadsheet. Clicking on File/Open opens the data input dialog box presented in Figure 18. Data is imported from an SPSS.sav format file as a flat file by clicking on Configure and browsing to the file.

Once the data is imported, each variable can be defined as an Input, a Target, or Reference and can be selected for use in the model. Neural Connections also provides the ability to analyze the distribution underlying each variable via the Input1 tool data sheet. Transformations can be tested in a straightforward manner, which a user can easily learn from the Neural Connection documentation.

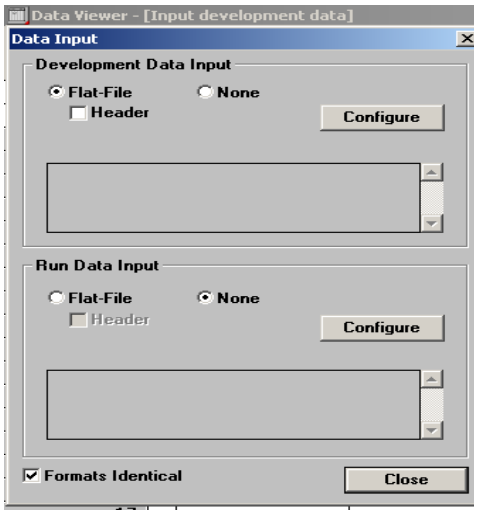


Figure 18. Neural Connections Data Input Window

Once the user is ready to set up the model, the first task is to allocate the data via the Input1 tool data sheet. Click on Data/Allocation to open the window presented in Figure 19. The options in this window are self-explanatory; however, the reader should note the two options for setting the seed value when a random process is selected. Random ordering can be applied to the data and/or the selection of the training and validation sets.

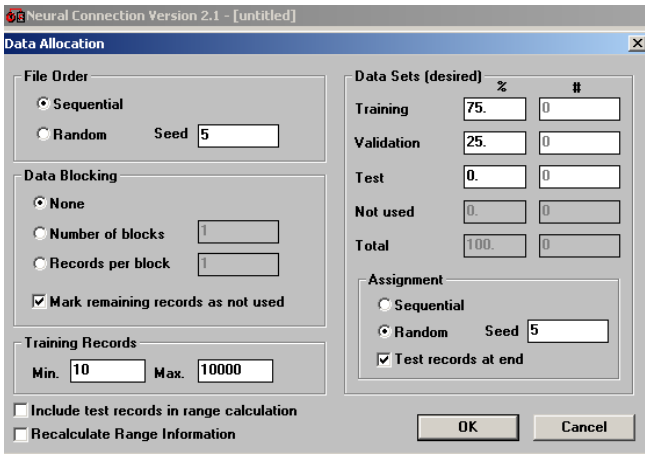


Figure 19. Neural Connection Input Data Allocation

Clicking on the tool the user has selected to define the network type (MLP in Figure 17 above), then clicking on dialog opens the window presented in Figure 20. This window provides the options for setting the network parameters. Note that the Conjugate Gradient method is the chosen network learning technique in this research. The Conjugate Gradient method is discussed in section 2.9.5. Network weights can be assigned randomly and the user has yet a third opportunity to change the seed number for this random process. Random seed numbers are changed manually for each model to provide robustness by training on different data sets and allowing different initial weights to avoid getting "stuck" in the same error surface local minima.

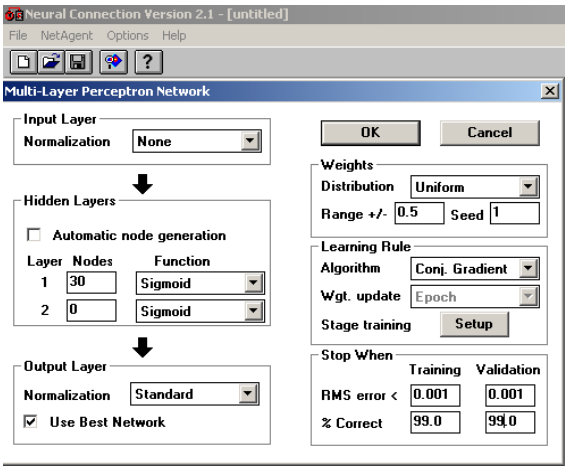


Figure 20. Neural Connection MLP Network Parameters

The final step to setting up the basic network architecture is to configure the output. Figure 21 presents the dialog box for Text1 in Figure 17 above. The dialog box in Figure 21 allows the user to select the data set for the trained model to be applied to and a delimited format for the text output. The path and output file name are also accomplished from this dialog box. The user should note that the size of data printed to a screen is limited and truncates the beginning of the output. Printing to a text file is the best approach.

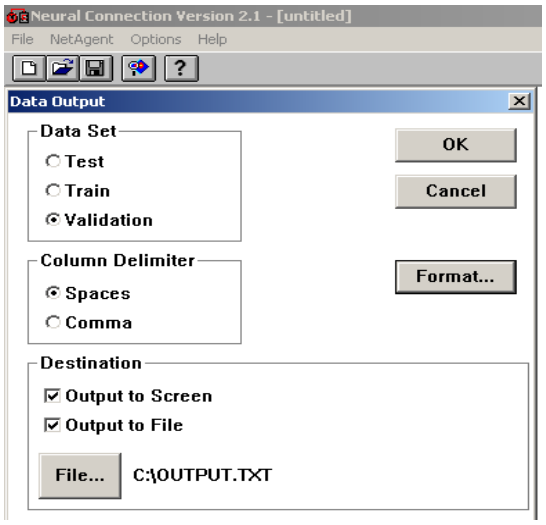


Figure 21. Neural Connections Output Dialog Box

Figure 22 presents the dialog box that opens when the Format button in Figure 21 is clicked. The options in this window are straightforward. Note that the Cross Tab Matrix option provides a confusion matrix. The number of bins selected defines the size of the confusion matrix. Two bins provide the usual 2 x 2 confusion matrix with the decision threshold set at 0.50. Selecting more bins for a two-group problem simply defines more evenly spaced decision thresholds. The output then provides an idea of how many network outputs of each group fall into the different confusion matrix bin ranges.

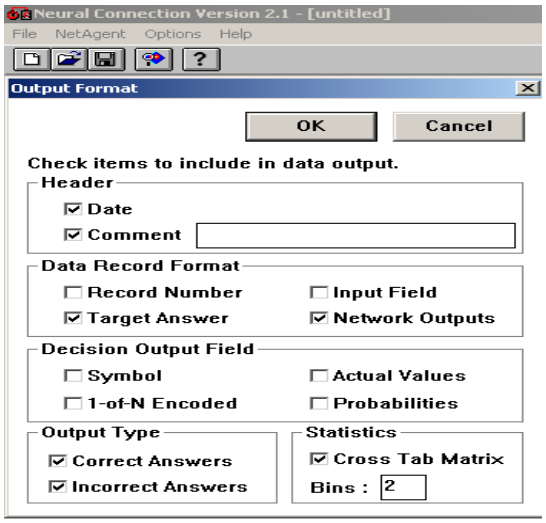


Figure 22. Neural Connections Output Format Dialog Box

Network weights can be viewed after training stops by clicking on the network tool, then clicking on the "Status" option in Figure 17. This brings up the window presented in Figure 23. Viewing the network weights in this way is convenient; however, capturing the weights for use in the SNR method is not straightforward. A special VBA application had to be written to capture and organize the network weights as they appear in Figure 23. Appendix C provides instructions on how to prepare the .NNI information, once it is opened in a text file, before the VBA application can be used. Appendix D provides the VBA code used to import and organize the network weights and input standardization parameters from a pre-processed text file of the .NNI network architecture into an Excel Spreadsheet.

Networks developed in Neural Connections 2.1 are saved as a .NNI file. The .NNI file is simply a text file containing the architecture of the network. The information presented quite neatly in Figure 23 is obtained by opening the .NNI file containing the network architecture from within a text file using the "Open" option from with in the

"File" menu and browsing to the appropriate .NNI file. The reader should consult the Neural Connections 2.1 documentation for an explanation of the information contained in the .NNI file.

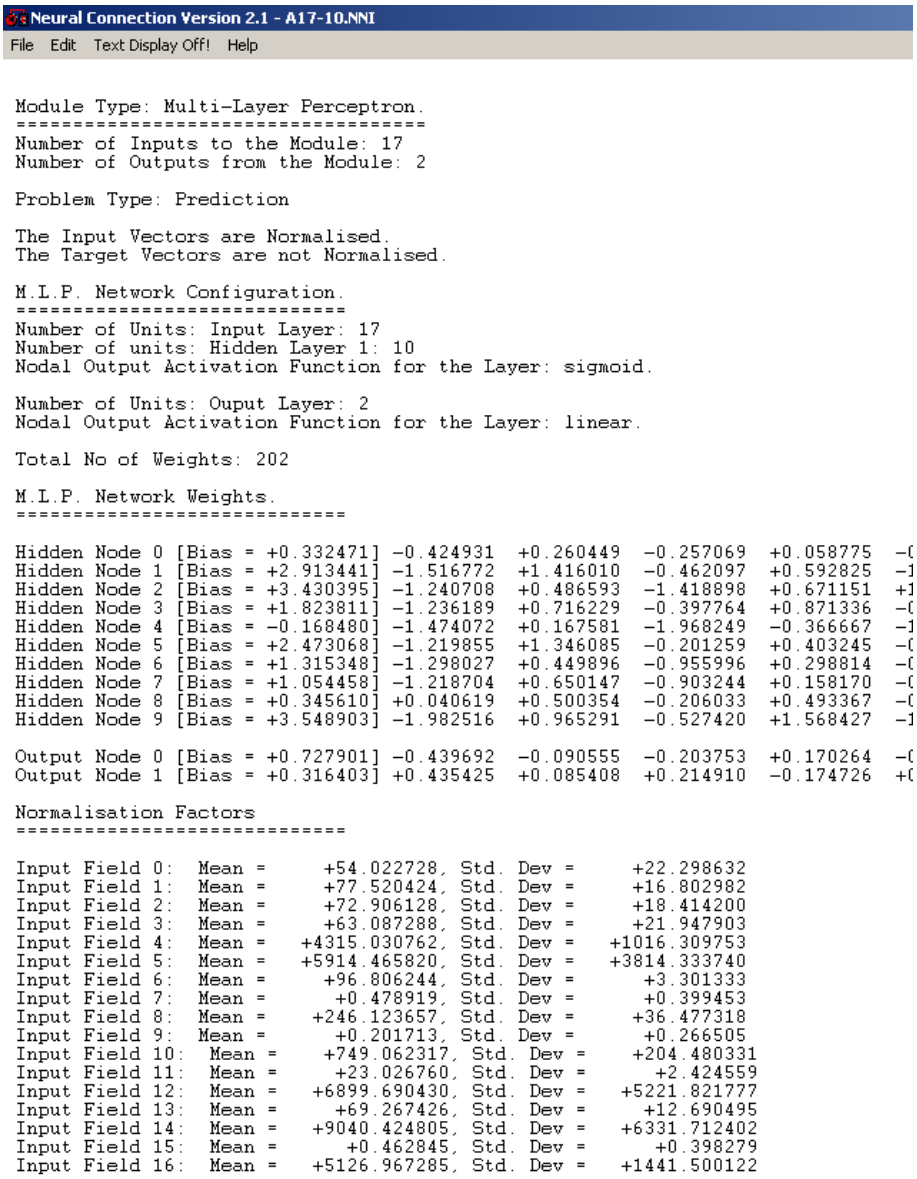


Figure 23. Network Weights and Standardization Parameters

3.5 Validation Study

The first objective of this research is to validate the current inputs to the PCSM model. This is done several ways. The first method involves a combination of linear regression and factor analysis. Variables retained by stepwise linear regression are compared to their associated factor analysis interpretation. Factor analysis is done in a confirmatory sense and is performed on the three data sets described in section 3.3. Factor analysis is performed using both unrotated factors and varimax rotation. Factor interpretation is performed using the resultant component factor loadings as a means of understanding the latent constructs underlying each factor.

Factor analysis was first performed using Kaiser's criterion, whereby the number of factors to use is defined by the number of correlation matrix eigenvalues that are greater than one. In each case, several factors beyond the ninth factor have eigenvalues very close to 1.0. Determining the number of factors via a scree plot of the eigenvalues dictated that the preferred number of factors to retain is eight. This held true for all three data sets. Figure 24 provides a sample scree plot of the eigenvalues associated with Data_A.

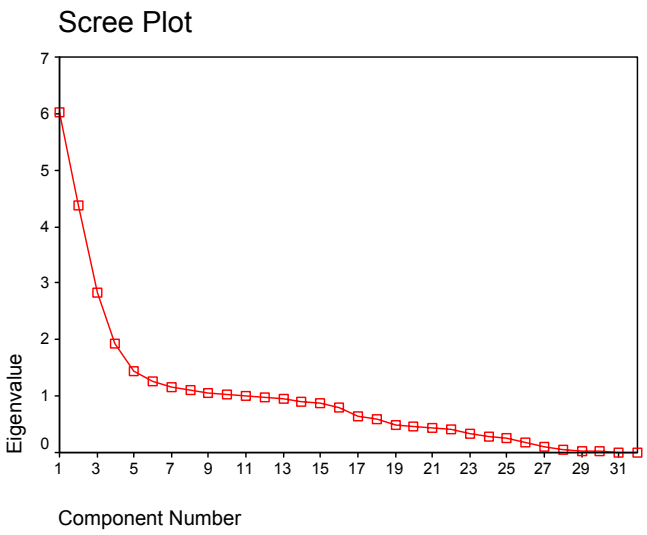


Figure 24. Scree Plot From Data_A

In this research, factor analysis helps overcome the situation where variables significant in a linear regression are sample dependent. For example, the 5 scores of the AFOQT are highly correlated and one or more may be present in the stepwise linear regression models. These variables have consistently strong factor loadings on their common factor. This is true for two reasons. First, the 5 AFOQT composites have many of the 16 AFOQT subtests scores in common, thus causing a degree of colinearity. The AFOQT is discussed in more detail in Chapter 2. Second, the AFOQT composites are measuring the same construct because the 16 AFOQT subtests primarily measure psychometric *g* (Earles & Ree, 1991) as well as factors commonly associated with paper & pencil tests such as verbal, quantitative, spatial, and perceptual speed (Skinner & Ree, 1987). Therefore, if another AFOQT composite takes the place of the Pilot composite, we can interpret the result as validating the underlying construct of the associated PCSM input rather than invalidating that input.

Stepwise linear regression is accomplished in SPSS as described in section 3.4. Probability of F-to-enter is set to 0.10 and probability of F-to-remove is set to 0.12. The resulting variables from each regression are reported in Chapter 4.

Partial Correlations are also used to validate PCSM via an iterative process of accounting for variables with the largest partial correlations. The correlations of all variables with the pass/fail criterion are first calculated. The largest correlation that is significant at the 0.10 alpha level is removed. Then the partial correlations are calculated with the removed variables' validity accounted for in the calculation. At each iteration, the most valid variable is removed and partial correlations calculated to account for all variables removed to that point. The process stops when correlations are no longer significant.

Finally, the effects of correcting correlations for range restriction are investigated. The correlations between the criterion and other variables known to be significant are first calculated. The variables are then ranked by the magnitude of the correlations. These correlations are then corrected for range restriction a new rank order is created. The order of the variables is then compared for uncorrected and corrected correlations to see if variable rankings change appreciably.

3.6 Regression Update

Recall that the DATA_A set used in the regression update is not the same as DATA_A used in the validation study because the TEST set is removed and the 10% attrition rate is required. DATA_B is not used in the regression update. Data_C is slightly changed also. A 50% attrition rate is considered unreasonably high for the

purposes of updating the regression weights. However, the influence of failure proportion on the resulting regression performance is of interest. Therefore, a the pass/fail proportions are set at 75%/25%, respectively. A version of DATA_C with 912 passes randomly selected from DATA_A give a 25% attrition rate (912 passes, 304 failures). Three different versions of this DATA_C are generated to guard against the effects of a "good" or "bad" sample on model performance. Unless otherwise noted, all results are for the TEST set.

3.6.1 Current PCSM Regression Discussion

The current PCSM model is regression based. The type regression type is unknown to the author. The results of this research suggest PCSM is based on linear regression rather than logistic regression. Based on information provided by AETC, the author had previously been under the impression that the actual regression weights were derived from a logistic regression and the discrete sigmoidal transform is simply a means of dealing with those raw PCSM scores greater than 1.0 as a result of the earlier linear regression. It appears that the discrete sigmoidal transform simply give the model a "logistic feel."

Some time after the original PCSM model was put into operational use, AETC chose to add a logistic aspect to the model by applying a discretized sigmoidal transformation to the linear regression output. This was done in order to address the problem of interpreting a PCSM score that ranged from zero to beyond 1.0 as a probability of passing UPT. In this research, both linear and logistic regressions are investigated.

Carretta (1991b) studied the effects of performing multiple linear regressions on criteria, which ranked observations according to simple weighted averages of T37 flying grades, T38 flying grades, and academic average. Due to non-normality and skewness among most of the inputs, Carretta (1992b) also performed regressions on Log transformed data. Carretta (1992b) reported negligible differences in results. The current data set suffers from the similar distributional issues. Based on Carretta's findings and in order to maintain consistency with the current PCSM model, the data is not transformed prior to performing the updated linear and logistic regressions.

Prior to performing the regression, several data preparation steps were required. The data provided by AETC includes the 9 raw BAT scores rather than the 5 actual BAT scores that are PCSM inputs. The process of transforming the 9 BAT scores into the required 5 BAT inputs presented in Figure 1 is presented again in Figure 25. The process is duplicated from a spreadsheet model provided by AETC.

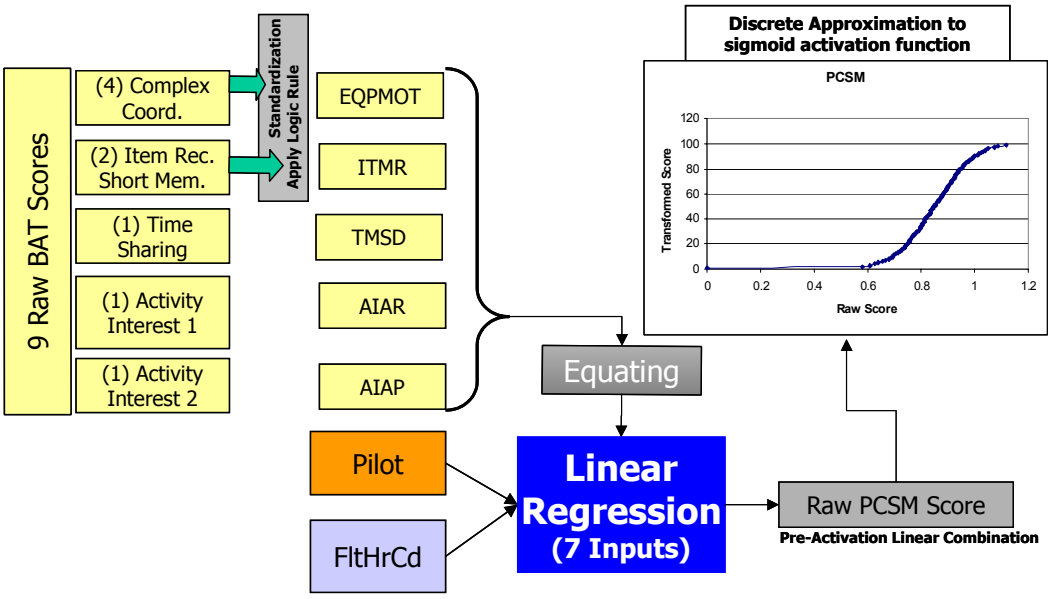


Figure 25. BAT Score Transformation Process

Four of the 9 raw BAT scores are 2-hand coordination pursuit tracking error scores. These are standardized, averaged, and multiplied by –1.0 to form a raw EQPMOT input. The raw EQPMOT scores are negated in order to prevent test compromise. Two of the 9 raw BAT scores relate to the reaction time and percentage correct for a short-term memory item recognition test. The raw input for PCSM is the reaction time, ITMR. In order to prevent test compromise, the ITMR input is scored at the maximal value of 2,500 milliseconds if the percentage correct, ITMP, is less than 75%. The final three raw inputs, TMSD, AIAR, AIAP, are not changed in anyway at this point. Each of these 5 scores are then transformed via equating tables. The equating tables date back to the original implementation of the operational form of PCSM. They equate the operational BAT test to the baseline BAT test configuration, which PCSM was developed on. Information concerning the equating tables can be found in Carretta &

Ree (1993a). The 5 resulting BAT scores are inputs into the current PCSM model along with the AFOQT Pilot composite and reported FAA flying hour code.

3.7 Independent Model

Three separate data sets are employed in the development of the independent model. The first, DATA_A, is the same data set used in the regression updates and consists of 2,732 passes and 304 failures. The second set, DATA_B, consists of equal numbers (N = 2,732) of passes and failures for a total of 5,464 records. Recall that bootstrap resampling with replacement was performed on the original 141 failures not included in the TEST set. The third data set, DATA_C, consists 141 passes and 141 failures. Unless otherwise noted, all results are for the same TEST set.

When networks are trained on any of the three data sets, the data can be randomly assigned to training and validation sets with proportions defined by the user in Neural Connections. In this research the training/validation proportions are set at 75% and 25%, respectively. For example, DATA_A contains 3,036 records (2,732 passes, 304 failures). Networks are trained on the 75% (N = 2,277) of records randomly assigned to the training set. As Neural Connections 2.1 trains the network, performance on the validation set is monitored. Training stops when validation set performance is optimized. Although use of a validation set prevents over training the network, model performance on the validation set will tend to be optimistic. The independent TEST set provides a more conservative estimate of model's generalized performance. The random training/validation set assignments are not changed for the networks trained to obtain a

mean SNR, but are changed at each iteration of he process feature selection from among the SNR ranked features.

3.7.1 Feature Selection and Network Development Algorithm

Figure 26 presents a flow diagram of the algorithm used to perform feature selection and develop and optimized multi-layer perceptron (MLP) neural network. Figure 26 is followed by a short description of each step of the algorithm. The methodology of the algorithm can be broken down into 4 main processes. (1) ranking features, (2) identify the minimum number of features to include, (3) optimize the number of hidden layer nodes, (4) reduce local minima effects.

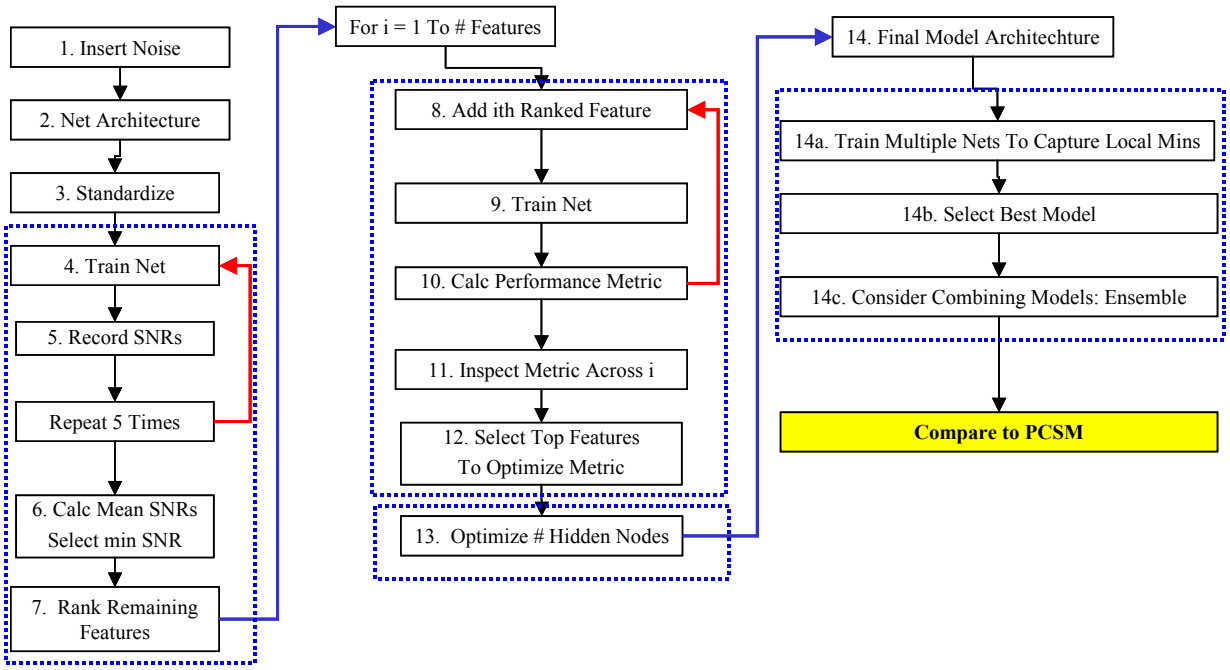


Figure 26. SNR Feature Selection & Network Optimization Algorithm

1. Introduce a Uniform (0,1) noise feature to the original set of features.

2. Standardize all features.

3. Randomly initialize the weights between 0.01 and 0.01.

4. Randomly select the training and validation sets.

5. Train the network.

6. Retain the weights that resulted in the best validation set classification error and calculate the SNR for each feature.

7. Store the SNR's for each feature. Repeat steps 3-7 for *n* runs. *n* is user defined and is data set specific. 5 runs were used in this research. Training/validation sets remained constant for this portion of the process.

8. Compute the mean saliency measure for each feature based on the *n* runs.

9. Rank order all features based on their mean saliency and remove all features whose mean saliency is less than some minimum mean saliency defined by the user after inspection of all mean saliencies. This is relative. SNR > 0 implies more saliency than random noise. If SNR's tend to be large, a feature with SNR < 3.0 may not be regarded as being salient.

10. Beginning with only the top ranked feature, iteratively train networks adding the next highest ranked feature one at a time until the sum of squared errors is optimized. Randomize the training/validation set assignments and initial weights with each iteration. This is done without a noise feature present. Neural Connections 2.1 automatically adds a bias term and its associated node in the hidden layer.

11. Retain the features present in the network with the optimal sum of squared error across the validation set.

12. Train additional nets with the retained features to ensure performance is not merely due to the random training/validation set assignment and/or the initial weights.

13. Optimize the number of nodes for a network with the retained features with a sigmoid activation function by training individual networks with different numbers of nodes in the hidden layer. Begin with a sufficiently small number of nodes and add one node iteratively until performance on the TEST set is optimized. This is done by inspection of ROC curves with the current network compared to the current PCSM model for the TEST set.

14. Investigate any networks with different numbers of nodes that perform similarly at optimal levels. Decide how many nodes to include in the final network architecture.

The signal-to-noise ratio (SNR) algorithm is used as the basis for performing feature selection in this research. Young (2002) modified the signal to noise algorithm as presented by Bauer et al. (2000). Young (2002) trained multiple networks and used the mean SNR as the saliency measure. Then, based on the mean SNR, the worst remaining features were removed one at a time until the classification accuracy was optimized.

In this research, 5 networks are trained to obtain a mean SNR for each of 34 features including the noise feature. At this juncture, each network contains one hidden layer with one node per network input. Neural Connections 2.1 automatically adds a bias input bringing the total number of inputs to 35. It is not necessary to account for this input when assigning the number of hidden layer nodes in the network dialog box. Data reduction is performed iteratively. It begins by training a network with the highest ranked input feature and calculating the sum of squared errors across the validation set. The next highest ranked feature is added and the process continues iteratively until the sum of squared errors is minimized.

Once the features are selected, the number of nodes in the hidden layer must be optimized. This is done by training single networks with a specified number of hidden nodes covering a wide range. The performance of each is investigated and candidates are further investigated to ensure performance is robust. With the network architecture defined, the final step is to train multiple networks with different initial weights to ensure the optimal network does not suffer from being "trapped" in a local minimum on the error surface.

3.7.2 Discriminant Analysis Feature Selection

In selecting features that provide the maximal separation between groups as defined by the Mahalanobis distance, it is necessary to consider higher order terms. In this research, a heuristic is used to search for significant quadratic and two-way interaction terms. SPSS performs stepwise discriminant analysis. Interaction terms are generated independently and provided to SPSS. This can be done within SPSS; however, it is more convenient to do in Excel.

Young (2002) investigated second-order terms for every variable available in the data set. SPSS was not able to perform stepwise discriminant analysis on such a large number of variables; therefore, Young (2002) performed a heuristic algorithm to investigate the interactions through a series of stepwise discriminant analyses. Each iteration performed stepwise discriminant analysis on the basic variables available and all second-order terms of one of those variables. In this research and to the extent possible due to differences in variable coding decisions, the final set of first and second-order terms that Young (2002) performed discriminant analysis on is replicated. A discriminant function is derived with these variables using the current data sets.

A smaller set of basic variables used to investigate second-order interactions was selected as a means of generating a second discriminant analysis model. 14 variables were selected for this purpose. These variables were selected based on two criteria. First, a stepwise linear regression was performed on DATA_A with 0.15 probability of F-ratio to enter and 0.20 to leave. This resulted in 13 significant variables. These 13 variables were compared to the 17 variables selected via the Signal-to-Noise ratio (SNR) method, which is discussed in section 3.7.1. After comparing these two sets of variables,

14 variables were selected as basic variables used to search for significant interaction terms. These three sets of variables are presented in chapter 4. Fortunately, SPSS is able to perform stepwise discriminant analysis on the set of 14 basic variables along with every second-order interaction term. Therefore, an algorithm similar to that used by Young (2002) is avoided.

3.7.4 Ensemble Method

In order to develop a final model, the ensemble method was used to combine individual networks. Three (3) networks with a single hidden layer of 23 neurodes were trained using the sigmoid, hyperbolic tangent, and linear activation functions. The hidden layer contained 23 neurodes as a result of steps 13 and 14 of process described in section 3.7.1. The ensemble method can have the greatest advantage when the combined networks find different local minima on the error surface (Perrone & Cooper, 1992). Initial network weights are selected from a random uniform (-0.5, 0.5) distribution. The increase in range of initial weights from +/- 0.01 to +/- 0.5 for these 9 networks allows training to begin from significantly different locations on the error surface. This is intended to maximize dissimilarity among the networks. Thus, correlation between the individual network errors should be minimized for several of the networks allowing the ensemble method to take advantage of different local minima.

The ensemble method can become saturated when it is applied to many networks (Perrone & Cooper, 1992). Hence, it is necessary to minimize the number of networks used to perform the ensemble method. With 9 trained networks it is likely that some networks are redundant (i.e. highly correlated network errors). Dissimilar networks are selected by compiling a matrix of the network errors in columns and performing factor

analysis with varimax rotation. Three (3) factors are retained and the highest loading network for each factor is selected for the ensemble method.

3.9 Model Comparison

The ROC curve is the primary means of comparing independent models to the current PCSM model. For this research, the ROC curve plots the probability of detecting a failure (true positive or target detection) versus the probability of falsely classifying a pass as a failure (false positive or false alarm). Equations 28 and 29 present these two probabilities as calculated from a confusion matrix that is based on a single decision threshold. The confusion matrix was discussed in Section 3.4.1.

$$\Pr ob(TP) = \frac{TP}{TP + FN} \quad \textbf{(27)}$$

$$\Pr ob(FP) = \frac{FP}{TP + FP} \quad \textbf{(28)}$$

Each point on a ROC curve represents these two probabilities for a given decision threshold. In this way, the performance of multiple models can be performed across the range of thresholds. Hence, a three dimensional performance is mapped into two dimensions. Furthermore, a ROC curve makes the trade off between maximizing True Positives and minimizing False Alarms intuitive. To illustrate how to compare two models using ROC curves, Figure 27 presents ROC curves for two notional models. Note that the bottom curve is wholly contained within the area of the top curve. This demonstrates a desirable condition known as stochastic dominance. A model is

stochastically dominant if it outperforms another model throughout the entire threshold range. In practice, two models often perform best in different intervals along the threshold range, thus overlapping of ROC curves is common.

In order include PCSM with model outputs that range from 0.0 to 1.0 on a ROC, it is necessary to transform the PCSM score from its range of 1-99 to 0.01-.99. This is done by multiplying PCSM scores by 0.01. SPSS forces unique variable names, so PCSM scores transformed in this way are named "PCSM2." Many of the ROC's presented in this research include "PCSM2" for performance comparisons.

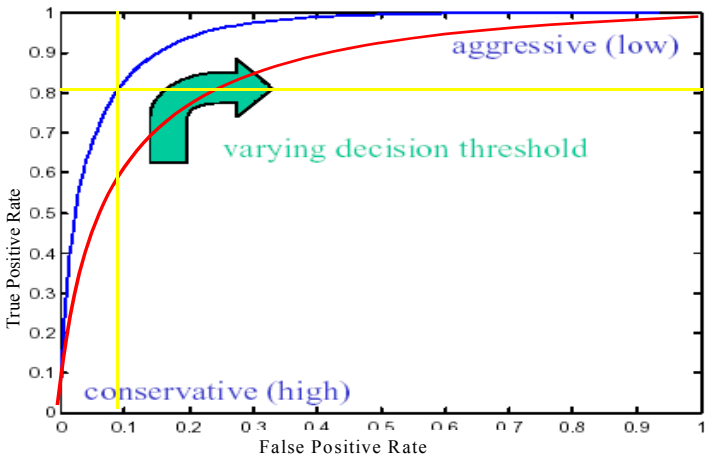


Figure 27. ROC Curves for Two Notional Models

3.10 Chapter Summary

This chapter provided the reader with an overview of the details regarding the most pertinent aspects of the analysis process used in this research. First, it reviewed the data preparation process required before analysis could begin. Next, it provides basic tutorials for the two specialized software packages used in this research. Thirdly, it

provides an overview of the methodology used to accomplish each objective of this research. In any multivariate analysis, one of the most crucial issues is feature selection. Two methods of feature selection are summarized, SNR and stepwise discriminant analysis. Finally, ROC curves are presented as the primary means of comparing the performance of competing models.

IV. Results

4.1 Introduction

As discussed previously, this research has three goals. They are validation of PCSM, updating the regression for the current PCSM model, and development of an independent model. The results in these three areas are presented in separate sections. Updated regression and independent model results are compared to the current PCSM model by way of the independent "TEST" set. Some results pertain to validation sets and are noted accordingly. Validation sets may be unique due to changing random number seeds within Neural Connections or manually reassigning training/validation sets via a uniform (0,1) number generator from within Excel. All references to the TEST set are the same set of 470 observations pulled out of DATA_A as described in Section 3.6.1.

4.2 Validation Study Results

Factor analysis was performed using both unrotated factors and varimax rotation on all three validation data sets in Table 13. For each data set, the first three factor interpretations are clearly in line with the three latent constructs underlying the current PCSM model inputs. These are the AFOQT test, the BAT test, and a measure flying experience. All factor interpretations and the variables underlying the assigned interpretation are presented in table 14

Table 13. Data Set Summary

	VALIDATION		LOGREG		INDEP-MODEL	
	PASS	FAIL	PASS	FAIL	PASS	FAIL
DATA_A	3,155	188	2,732	304	2,732	304
	96%	4%	90%	10%	90%	10%
DATA_B	3,155	3,155	NA	NA	2,732	2,732
	50%	50%	NA	NA	50%	50%
DATA_C	188	188	912*	304	141	141
	50%	50%	75%	25%	50%	50%
TEST	NA	NA	423	47	423	47
	NA	NA	90%	10%	90%	10%

*** 3 of these data sets were created**

The factor interpretations remained consistent across all three data sets for both unrotated and varimax rotated factors. Varimax rotation made the interpretations much more apparent. There is very little ambiguity in the varimax rotated factor loadings matrix. The factor number associated with a specific factor interpretation did tend to change across the three data sets to a small degree. For example two successive factors might swap order. The variables loaded consistently on their respective factors regardless of the factor number, thus interpretation across all three data sets was made using the same variables.

Table 14. Factor Analysis Interpretations

Factors	1	2	3	4	5	6	7	8
Interp.	AFOQT	BAT Complex Coord.	Basic Fly Exp	Age	Adv. Fly Exp	BAT Factor 1	BAT Factor 2	Looks On Paper
Variables	Pilot Nav Quant Verbal Acad	BAT H2CX1 PS2X2 PS2Y2 PS2Z2	FltHrCd AeroCd	BAT_Age ROTC AD	AeroCd Multi Instrum Instruct	ITMR AIAR	AIAP	GPA OTS_AD OTS_Civ AFR ANG

Note that the second factor is defined by the BAT test. Specifically, it is the EQPMOT related BAT scores that define this factor. Regression results in sections 4.3.2 and 4.3.3 and PCSM input rankings in section 4.3.4 indicate that the EQMPOT is the most significant of the BAT inputs to PCSM.

The last factor interpretation is interesting to note. There seems to be a latent construct explained by variables associated with GPA and UPT candidates selected from sources other than the AFA or ROTC. Given the fact that the AFA and ROTC supply the majority of UPT candidates, competition for available slots at the other 5 selection sources is great. I chose to interpret the latent construct as an input that describes the ability of those selected from these pools to "look good on paper." AD loaded moderately high on this factor as well, but had a much higher loading on the age factor. Nevertheless, AD could also be used as part of the interpretation of factor 8. Interestingly, the age factor resulted in strong loadings for both ROTC and AD, but were opposite in sign. This is intuitive given that AD applicants are generally older than ROTC cadets. Although only moderate in magnitude when compared to other loadings

for factor 5, non-AFA and non-ROTC source indicators suggest that advanced flying experience is a significant to selection from those sources. The varimax rotated factor loadings for DATA_A are presented in Appendix F.

Stepwise linear regression was performed on each data set with the pass/fail criterion as the dependent variable. Table 15 provides a summary of these results. Significant variables are preceded by the factor number, for which the variable is used in factor interpretations presented in Table 14. The probability of F-ratio to enter is set at 0.10 allow more variables to enter the regression, but removal is set to 0.12 to remove variables quickly as they become insignificant. This is done to investigate variables with predictive power rather than sheer parsimony. Note that for the bootstrapped data set, the probabilities to enter/leave were set at 0.05/0.10 to limit the number of variables that become significant with such a large sample size. Although adjusted R-squared values are not very impressive, regression significance is great. This is due to large sample sizes, which causes relatively small MSE results and allows for larger F-ratios.

Table 15. Stepwise Linear Regression Results

Data Set	95% / 5%		Bootstrap 50% / 50%		Small 50% / 50%	
Enter/Leave	0.10	0.12	0.05	0.10	0.10	0.12
Vars (Factor)	(1) Pilot (3) Aero-Cd (2) PS2Z2 (2) TMSD (4) ROTC (4) BAT_Age (4) AD		(1) Pilot (3) Flt Hr Cd (2) PS2Z2 (2) TMSD (4) BAT_Age (4) ROTC (1) Verbal (7) AIAP (4) AD	(3) Single (5) Multi (1) Acad (2) BAT (2) PS2X2 (5) Instrum (1) Nav (2) HC2X1 (5) Instruct	(1) Pilot (3) Flt Hr Cd (2) BAT (4) BAT_Age	
R	0.220		0.475		0.441	
Adj R-sqd	0.027		0.223		0.186	
SE	0.225		0.441		0.452	
F-Ratio	24.32		76.25		22.38	
MSE	0.051		0.194		0.204	
SSE DoF	3,335		4,713		371	

The variables are also investigated for validity of predicting a binary pass/fail criterion. Calculating correlation coefficients and selecting the variable with the largest significant correlation at the 0.10 alpha-level begins this process. Once a variable is selected, partial correlation calculations are made by "partialing out" all variables previously selected. This is done iteratively using SPSS's partial correlation capability. Table 16 provides a list of the variables in the order in which they were selected. The first 7 variables had significant partial correlations at the 0.10 alpha level. The last 6 partial correlations where selected at successive iterations on the basis of partial correlation magnitude alone. For example, the IFT variable has the largest partial correlation of the last 6 non-significant variables selected, but it was selected last. At iteration 12 the partial correlation for IFT was 0.011 with a p-value of 0.531.

Table 16. Partial Correlation Results

Order	Variable	Partial Corr	p-value	Order	Variable	Partial Corr	p-value
1	PCSM	0.184	0.000	**8**	Verbal	-0.24	0.174
2	Pilot	0.075	0.000	**9**	Quant	0.018	0.304
3	AD	-0.033	0.054	**10**	Nav	-0.028	0.101
4	ROTC	-0.045	0.010	**11**	TMSD	0.020	0.238
5	BAT_Age	-0.041	0.017	**12**	Flt Hr Cd	0.022	0.196
6	PS2Z2	-0.036	0.039	**13**	IFT	0.031	0.077
7	AIAP	0.033	0.060				

The effects of correcting correlations for range restrictions were also investigated. Table 17 presents both the uncorrected and corrected correlations between the pass/fail criterion and a set of 10 variables related to the current PCSM model. The variables are ranked by magnitude of correlation. Note that the rank order of the variables changes little. This suggests that in terms of variables used in the current PCSM model, the affects of range restriction are not that great. PCSM remains the single most predictive variable, followed by variables that are actual PCSM inputs. Although, the AFOQT Navigator composite is not in the current PCSM model, it is highly correlated with the Pilot composite.

Table 17. Correlations Corrected for Range Restriction

Correlations with the Pass/Fail Criterion			
Variable	Uncorrected	Variable	Corrected
pcsm	0.189	pcsm	0.219
pilot	0.180	pilot	0.201
flthrcd	0.143	nav	0.169
nav	0.128	flthrcd	0.154
ps2z2	-0.089	ps2z2	-0.131
hc2x1	-0.083	bat	0.119
bat	0.080	hc2x1	-0.112
tmsd	0.080	itmr	-0.068
itmr	-0.047	aiap	0.050
aiap	0.033	tmsd	0.028

4.3 Regression Update Results

The regression weights are update with linear and logistic regressions. Both are compared to the current PCSM model. AETC/SAS provided an Excel spreadsheet, which is currently used to perform this pre-processing and apply the regression weights. The regression weights were updated using the same pre-processing steps currently employed to transform the 9 raw BAT scores into the 5 BAT scores that actually enter PCSM. The other two inputs, AFOQT Pilot and Flight Hour Code are not transformed prior to applying the regression weights. This pre-processing was graphically illustrated in Figure 25.

Unfortunately, the 5 transformed BAT scores are not recorded by AETC/SAS. Therefore, an Excel spreadsheet tool was created to apply the standardization, averaging, and other logical operator steps of the AETC/SAS Excel tool to the entire DATA_A and TEST sets. MatLab is then employed to apply the equating tables for the 5 BAT inputs to the entire data set. AETC currently applies the equating tables manually on a record-by-record basis in a PCSM calculator Excel spreadsheet. The MatLab code for the EQPMOT input is provided in Appendix B.

Once the raw data provided for this research is transformed into the actual PCSM inputs, the regression can be modeled. Currently, AETC/SAS calculates the linear combination of the weights and inputs to get a raw PCSM score. Then a discrete approximation to an unknown form of a sigmoidal activation function is applied. Figure 28 presents the approximate activation function created from the look-up table currently employed by AETC.

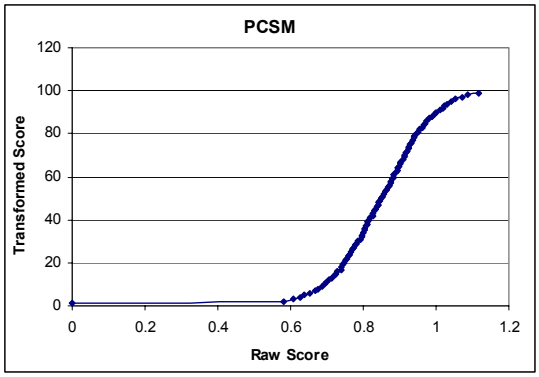


Figure 28. Current Activation Function For Raw PCSM Scores

The lookup table used to create Figure 28 has a maximum raw score value of 1.116. All raw PCSM scores greater than 1.116 are assigned a PCSM score of 99. In this data set, the raw PCSM scores range from 0.0 to 1.24. This is a consequence of linear regression. Logistic regression outputs are bounded by 0.0 and 1.0. This introduced the issue of how to treat the updated logistic regression output since it has already passed through a continuous sigmoidal activation function. Although the sigmoid activation function has already been applied to the updated logistic regression output, AETC/SAS prefers to apply the current approximation function in Figure 28. This is accomplished

by rescaling the logistic regression output to exhibit a range of 0.0 to 1.116. The discrete activation function in Figure 28 is then applied.

4.3.1 Logistic Regression Results

Performance results of the updated logistic regression on the DATA_A set are similar to the current PCSM model when applied to the TEST set. This held for both the raw logistic regression output as well as the transformed output. Figure 29 shows the performance of the raw updated logistic regression output. Figure 30 shows the performance of the updated logistic regression after the output range was changed to (0, 1.116) and the approximate activation function in Figure 28 was applied.

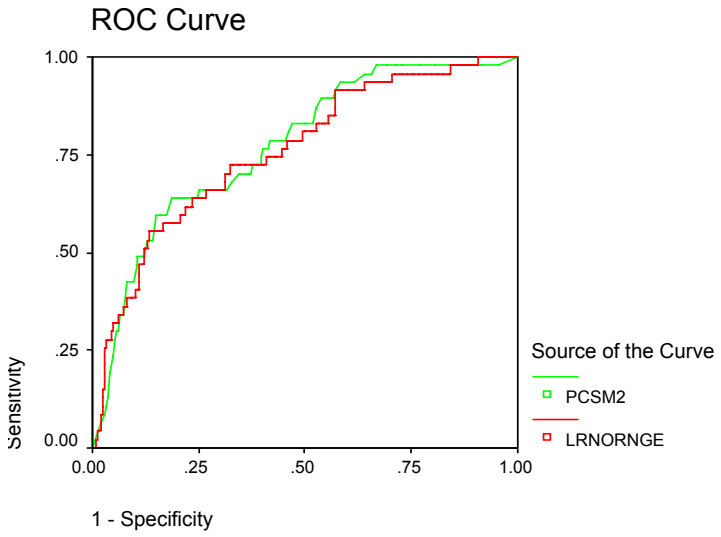


Figure 29. Raw Logistic Regression vs PCSM

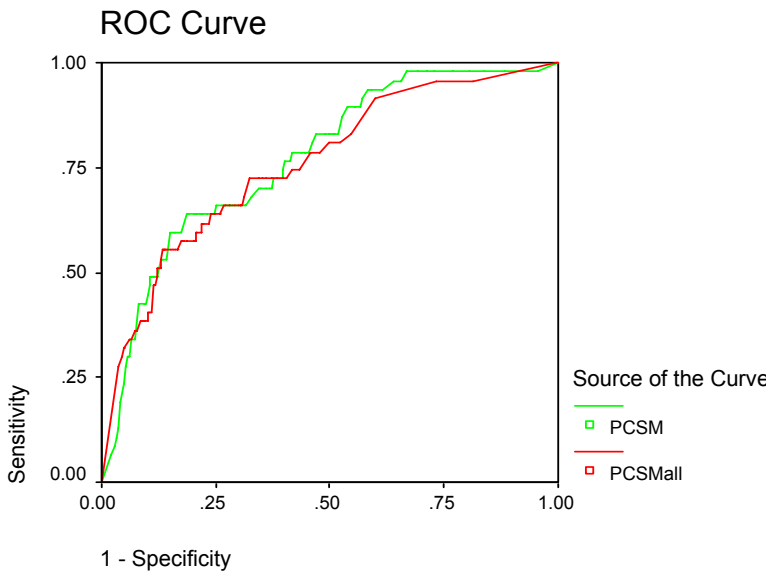


Figure 30. Range Converted & Current Activation Applied vs PCSM

As discussed in section 3.6.1, a logistic regression was also calculated for DATA_C1, DATA_C2, and DATA_C3. These data sets consisted of 75% passes (N = 912) and 25% bootstrapped failures (N = 304). These 3 logistic regressions resulted in performance similar to DATA_A. An example is provided in Figure 31. These results are presented for raw logistic regression output. Again, rescaling and applying the approximate sigmoidal function results in quite similar performance. Comparing the results in Figures 29 and 31, it appears that performance of the logistic regression is not sensitive to the proportion of failures in the data set (10% vs. 25%). It appears that updating the current PCSM regression with a truly logistic regression does not appreciably improve the model for the current data set. These results also suggest that forcing the use of the approximate sigmoid function on logistic regression results is not necessary. These results provide further validation of the current PCSM model.

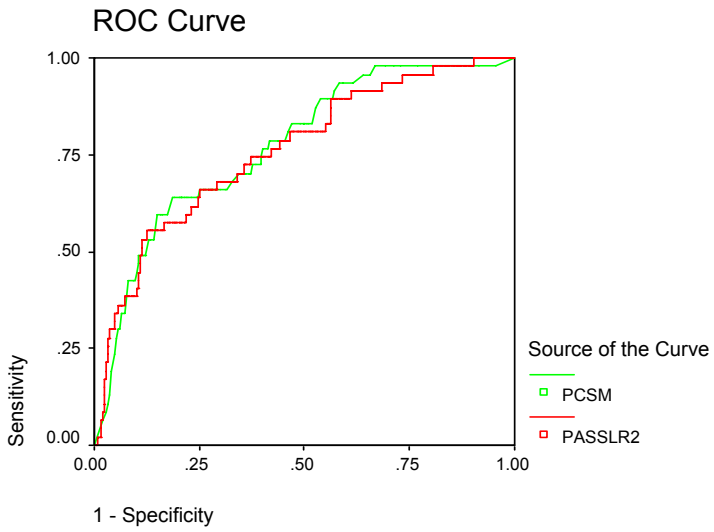


Figure 31. Performance of DATA_C2 Raw Outputs vs PCSM

Analysis of the logistic regression results would not be complete without inspection of the resulting weights and their associated significance. Table 18 provides a summary of the current PCSM regression weights and the logistic regression weights derived on DATA_A. Table 19 provides the same PCSM weights and logistic regression weights for 3 separate DATA_C sets. The p-values reported are for the logistic regressions. It is interesting that although performance for all four logistic regressions is similar to the current PCSM model, the intercept and regression weights are substantially different.

Table 18. Logistic Regression Weight Summary for DATA_A

	PCSM	DATA_A	
Input	**Wt**	**Wt**	**p-val**
Intercept	.68966	-1.886	.006
Pilot	.00293	.0310	<.001
FltHrCd	.02313	.1316	<.001
EQPMOT	.08915	.2896	.019
ITMR	.00090	-.00002	.935
TMSD	.00010	.0029	.146
AIAP	.00240	.0111	.038
AIAR	.00004	-.00003	.562

Table 19. Logistic Regression Weight Summary for 3 DATA_C Sets

	PCSM	DATA_C1		DATA_C2		DATA_C3	
Input	**Wt**	**Wt**	**p-val**	**Wt**	**p-val**	**Wt**	**p-val**
Intercept	.68966	-4.063	<.001	-2.325	.004	-2.735	<.001
Pilot	.00293	.0325	<.001	.0286	<.001	.0279	<.001
FltHrCd	.02313	.1133	<.001	.1428	<.001	.1384	<.001
EQPMOT	.08915	.4415	.002	.2595	.061	.2472	.075
ITMR	.00090	.0002	.606	-.0002	.517	.0002	.554
TMSD	.00010	.0038	.092	.0022	.342	.0033	.141
AIAP	.00240	.0158	.011	.0092	.127	.0092	.127
AIAR	.00004	.00004	.590	-.00005	.442	-.00007	.319

4.3.2 Linear Regression Results

The weights resulting from linear regressions are much more similar to the current PCSM model than were the logistic regression weights. The weights for all four linear

regressions are presented in Table 20 along with the current PCSM weights. The similarity in intercepts for PCSM and the DATA_A regression suggests that DATA_A is similar to the data set used to in deriving the current PCSM regression weights. In this case, similar performance would not be as surprising as it is for the logistic regressions. Performance of the linear regression for DATA_A applied to the TEST set is presented in Figure 32. The performance of the linear regressions for the 3 data sets with 75% passes and 25% failures is similar the that in Figure 32. Figure 33 presents a comparison of the raw outputs for the linear and logistic regressions for DATA_A applied to the TEST set. The results are strikingly similar and suggest that the only advantage offered by a logistic regression is in its more realistic probabilistic interpretation.

Table 20. Linear Regression Weights

Linear Regression Results

	PCSM	DATA_A	DATA_C1	DATA_C2	DATA_C3
Intercept	0.68966	0.49506	-0.17864	0.12752	0.04433
PILOT	0.00293	0.00316	0.00606	0.00533	0.00521
FLTHR_CI	0.02313	0.00988	0.01694	0.02261	0.02191
EQP	0.08915	0.03181	0.07966	0.04695	0.04670
ITMRT	0.00090	0.00000	0.00003	-0.00004	0.00004
TMS	0.00010	0.00024	0.00066	0.00037	0.00055
AIAPER	0.00240	0.00079	0.00229	0.00122	0.00142
AIART	0.00004	0.00000	0.00001	-0.00001	-0.00001

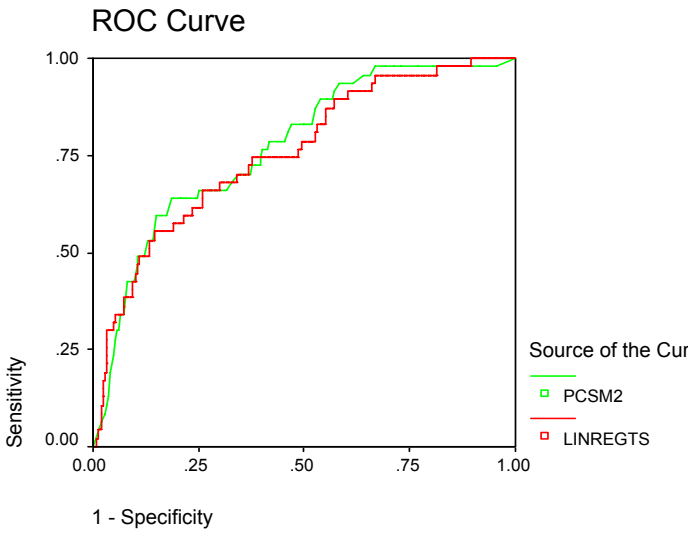


Figure 32. DATA_A Linear Regression Performance on the TEST Set

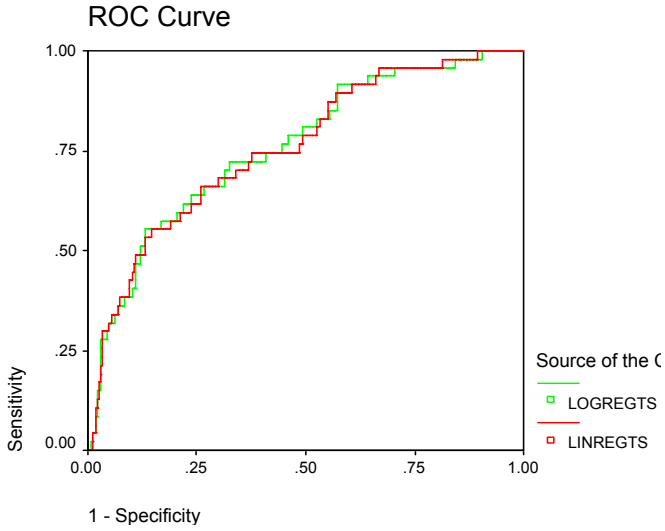


Figure 33. Comparison of Linear and Logistic Regressions on the TEST Set

4.3.3 Linear Regression Results for Updated EQPMOT Standardization

The current PCSM model combines 4 pursuit tracking error scores from the BAT. These 4 scores are standardized before being averaged and multiplied by –1.0. The

previous regression results used the same means and standard deviations as the current PCSM model. Results presented here reflect updated standardizations using the means and standard deviations observed in DATA_A. Figures 34 and 35 present the results of these linear regressions when applied to the TEST set. Figure 35 applies the approximate sigmoidal function applied in the current PCSM model. Again, application of the sigmoidal function does not improve performance. Updating the EQPMOT standardization results in performance degradation. This and the previous results suggest that no changes in PCSM's regression weights or the standardization process are warranted.

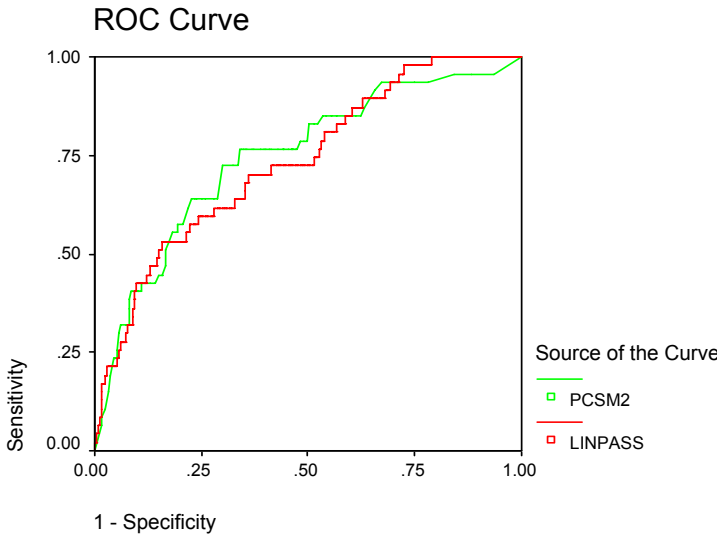


Figure 34. Linear Regression with Updated Standardization

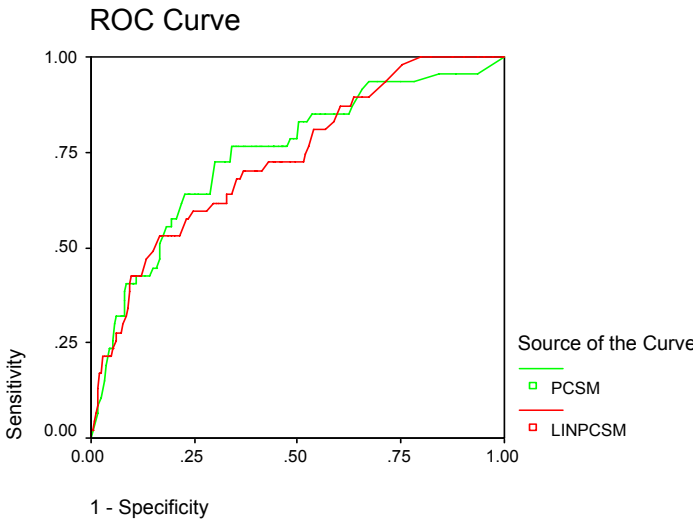


Figure 35. Updated Standardization with Approximate Sigmoid Applied

4.3.4 Investigating the Current PCSM Model

PCSM's continued positive performance initiated an interest in better understanding the PCSM model itself. With PCSM score as the dependent variable, linear regressions were performed on the current PCSM inputs with the current and updated EQPMOT standardizations. The regressions are performed on all available data. This includes both DATA_A and the TEST set combined. The regression parameters are presented in Tables 21 and 22. Adjusted R-Squared values for the two regressions are 0.942 and 0.940, respectively. The standardized coefficients match very closely across the two standardization schemes. Hence, the inputs are providing similar information in predicting the PCSM score with both standardizations. This indicates that the input distributions for the current data must be similar to those of the data used to derive the current PCSM model. Such similarities could explain why the current regression performance is much like the current PCSM model.

Table 21. Regression on PCSM with Current Standardization

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.650	1.462		7.285	.000
	PILOT	.660	.009	.353	77.111	.000
	FLTHR_CD	5.045	.039	.576	129.077	.000
	EQP	17.349	.286	.280	60.735	.000
	ITMR	-.020	.001	-.139	-32.519	.000
	TMSD	.177	.004	.190	42.491	.000
	AIAP	-.457	.011	-.175	-42.131	.000
	AIAR	-.007	.000	-.239	-56.464	.000

a. Dependent Variable: PCSM

Table 22. Regression of PCSM with the Updated Standardization

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	19.798	1.541		12.844	0.000
	PILOT	0.652	0.009	0.347	74.376	0.000
	FLTHRCD	5.034	0.040	0.573	126.459	0.000
	EQPMOT	12.452	0.209	0.286	59.629	0.000
	ITMR	-0.020	0.001	-0.138	-31.946	0.000
	TMSD	0.165	0.004	0.177	38.274	0.000
	AIAP	-0.455	0.011	-0.172	-40.957	0.000
	AIAR	-0.007	0.000	-0.240	-56.008	0.000

a. Dependent Variable: PCSM

Table 23 summarized the ranking of the PCSM inputs based on predicting Pass/Fail and PCSM itself. Both the current and updated EQPMOT standardization are represented for each dependent variable. The SNR rankings resulted from training 8 MLP networks with PCSM as the target and the current PCSM inputs as input features. Table 24 presents the mean SNR ratios for these networks. It is interesting to note that the Pilot and flight hour code rankings are reversed when PCSM is the dependent

variable. This indicates that the flight hour code input is more important than the Pilot composite in determining an applicant's PCSM score, but the Pilot composite is more important for predicting the pass/fail criterion.

It is also worth noting that the top three inputs are always Pilot, FltHrCd, and EQPMOT. Among the 5 BAT inputs to PCSM, EQPMOT consistently proves to be the dominant predictor. Recalling the factor analysis interpretations provided Section 4.2, one can easily see the continued validation of the first three factors representing the PCSM inputs. Despite the fact that the Pilot composite is most predictive of UPT performance, the results in Table 23 suggests that FltHrCd is more predictive of the PCSM score. This discrepancy should be addressed in the PCSM model.

Table 23. Summary of Input Ranks Across Multiple Regressions

	Pass/Fail as Criterion		PCSM as Criterion		
→ EQPMOT Std	Old LinReg	New LinReg	Old LinReg	New LinReg	Old SNR
1	Pilot	Pilot	FltHrCd	FltHrCd	FltHrCd
2	FltHrCd	FltHrCd	Pilot	Pilot	Pilot
3	EQPMOT	EQPMOT	EQPMOT	EQPMOT	EQPMOT
4	AIA%HR	TMSAVDIF	AIART	AIART	AIART
5	TMSAVDIF	ITMRT	TMSAVDIF	TMSAVDIF	AIA%HR
6	AIART	AIA%HR	AIA%HR	AIA%HR	TMSAVDIF
7	ITMRT	AIART	ITMRT	ITMRT	ITMRT

Table 24. Mean SNR's from 8 Networks with PCSM as Target

Input	Mean SNR	Std Dev	90% LCL	90% UCL
FltHrCd	25.24	5.87	21.30142	29.18631
PILOT	20.81	5.90	16.84871	24.77144
EQP	19.97	5.75	16.11209	23.83586
AIART	18.67	5.81	14.76936	22.5686
AIA%	16.41	5.75	12.54278	20.26943
TMS	16.35	6.09	12.2556	20.44272
ITMRT	13.77	5.83	9.855359	17.68747

In order to more fully investigate the relationship between PCSM and its inputs, linear regressions were performed on a series of sub-sets of PCSM inputs. The pass/fail criterion is the dependent variable in these regressions. First all 7 inputs are included, then the lowest ranked PCSM input was removed and a new regression performed. This was done repeatedly until only the Pilot input remains. The variables were removed according to the rankings of the linear regression results using the "old" standardization under the pass/fail criterion in Table 23. Hence, ITMR is the first input removed, followed by AIAR, TMSD, AIAP, EQPMOT, and FltHrCd. Table 25 presents the correlations among the linear regression outputs. The correlations above the diagonal represent the correlations among the outputs for DATA_A, while the correlations below are for the TEST set. PCSM-1 represents the model with the lowest ranked input removed. Likewise, PCSM-6 represents the model with the 6 lowest ranked inputs removed. Inspection of Table 25 reveals that correlations between regression outputs for the current PCSM model and the sub-models remain high until only the Pilot and FltHrCd inputs remain.

Table 25. Correlations Among Regression Outputs for PCSM Sub-Models

Linear Regression Output Correlations: Upper => DATA_A , Lower => TEST							
	PCSM	**PCSM-1**	**PCSM-2**	**PCSM-3**	**PCSM-4**	**PCSM-5**	**PCSM-6**
PCSM	1	0.867	0.859	0.848	0.872	0.811	0.671
PCSM-1	0.890	1	1.000	0.996	0.988	0.962	0.875
PCSM-2	0.884	1.000	1	0.996	0.989	0.962	0.876
PCSM-3	0.870	0.996	0.996	1	0.993	0.966	0.879
PCSM-4	0.890	0.989	0.990	0.993	1	0.973	0.886
PCSM-5	0.840	0.964	0.964	0.968	0.976	1	0.910
PCSM-6	0.713	0.877	0.876	0.879	0.889	0.913	1

Figures 37, 38, and 39 present the performance of the last three linear regressions, PCSM-4, PCSM-5, and PCSM-6. These figures present performance on the TEST set. Figure 37 presents PCSM-4, which includes Pilot, FltHrCd, and EQPMOT. Figure 38 presents PCSM-5, which includes Pilot and FltHrCd. Finally, Figure 39 presents PCSM-6, which includes only the Pilot input. The performance of the linear regressions for PCSM-1, PCSM-2, and PCSM-3 are all nearly identical to PCSM-4. Figure 37 shows that performance begins to degrade when EQPMOT is removed. Performance is worst when only the Pilot input remains. These results suggest that a more parsimonious model may be obtained through the use of top 3 ranked PCSM inputs; Pilot, FltHrCd, and EQPMOT. One would need to further investigate the importance of the other PCSM inputs before eliminating them from the model.

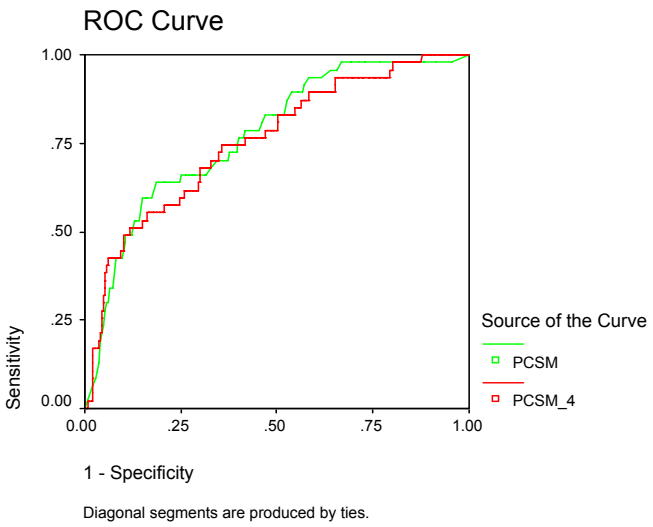


Figure 36. PCSM vs. PCSM, FltHrCd, EQPMOT

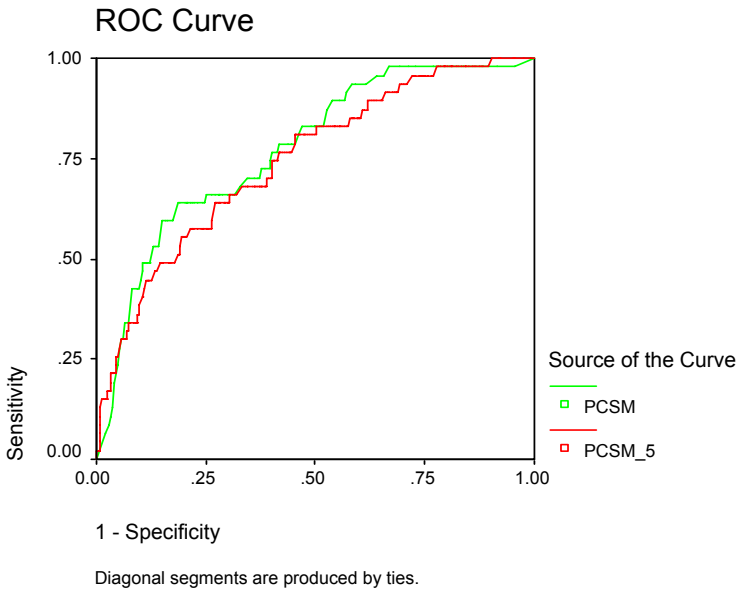


Figure 37. PCSM vs. PCSM FltHrCd

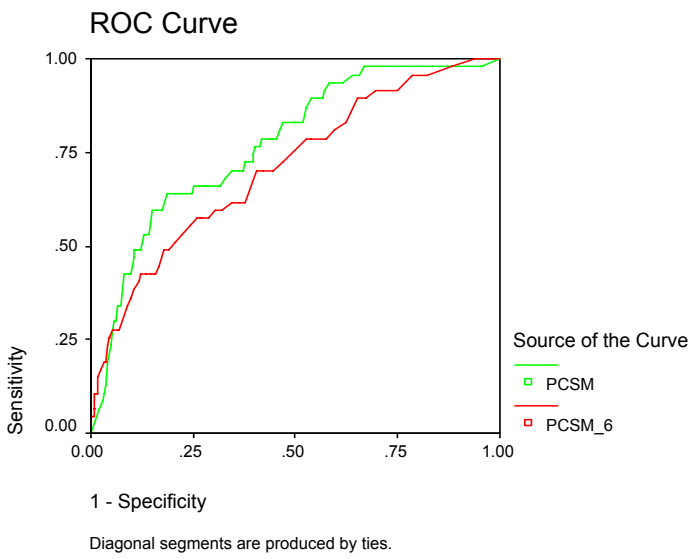


Figure 38. PCSM vs. Pilot

4.4 Independent Model Results

4.4.1 Signal to Noise Ratio Feature Selection Results

PCSM is considered a valid feature for use in developing an independent model. However, it is not considered for inclusion here because the ultimate goal of an improved model would be to replace PCSM. Including PCSM in an independent model would simply add complexity to an already complex model. With that said, the Signal-to-Noise (SNR) method was used to investigate feature saliency with and without PCSM included as a feature. This was done with the goal of understanding which features PCSM accounts for and which ones provide additional predictive information beyond PCSM.

Features with higher mean saliencies than PCSM are more important to the networks than PCSM itself. This would lead one to conclude that the importance of such features will not decrease upon removal of PCSM from the network. Table 26 shows this

to be the case. Table 26 presents the top 17 candidate features along with their mean saliency. It is interesting to note that PCSM ranks ninth among the features when it is included. A complete table of features and their mean saliencies is presented in Appendix E.

Upon PCSM's removal, features that become salient could provide additional insight into the question, "What makes PCSM work?". In fact, inspection of the complete list of mean saliencies in Appendix E shows that with PCSM removed, every available variable has a mean saliency of at least 3.0. This indicates that although PCSM is not the most important input for a neural network, it does account for much of the information available in the rest of the variables. This is also indicated by the fact that mean SNR's increase for every feature when PCSM is removed.

Table 26. Mean SNR Saliency For Feature Selection

PCSM Included (N = 8)			**PCSM Excluded (N = 5)**		
Rank	**Feature**	**Mean SNR**	**Rank**	**Feature**	**Mean SNR**
1	Pilot	15.05	**1**	BAT	21.14
2	BAT	14.62	**2**	Pilot	21.07
3	ROTC	12.02	**3**	Nav	17.39
4	Quant	11.47	**4**	Quant	15.91
5	AERO34	10.88	**5**	AIAR	15.49
6	ITMR	10.46	**6**	PS2Z2	15.26
7	BAT_Age	10.21	**7**	ITMP	15.23
8	PS2Z2	10.14	**8**	ROTC	14.78
9	**PCSM**	9.86	**9**	TMSD	14.12
10	AERO2	9.63	**10**	AERO34	14.00
11	H2CX1	9.21	**11**	ITMR	13.95
12	Instruct	8.97	**12**	BAT_Age	13.38
13	TMSD	8.84	**13**	PS2Y2	12.65
14	PS2Y2	7.94	**14**	AIAP	12.29
15	Nav	7.65	**15**	PS2X2	11.83
16	AIAR	7.51	**16**	AERO2	11.81
17	ITMP	7.46	**17**	H2CX1	11.79

With the features ranked by mean SNR, it becomes necessary to identify a parsimonious set of features for use in the final model. This is done in a forward selection manner by beginning with the highest ranked feature, training an individual network, and calculating the SSE across the observations in the validation set. Next, the second highest ranked feature is added to the network, training is performed, and SSE recalculated. Figure 39 presents SSE across the validation set for each network as this iterative process progressed. At this point in the research, the training and validation sets are being held constant.

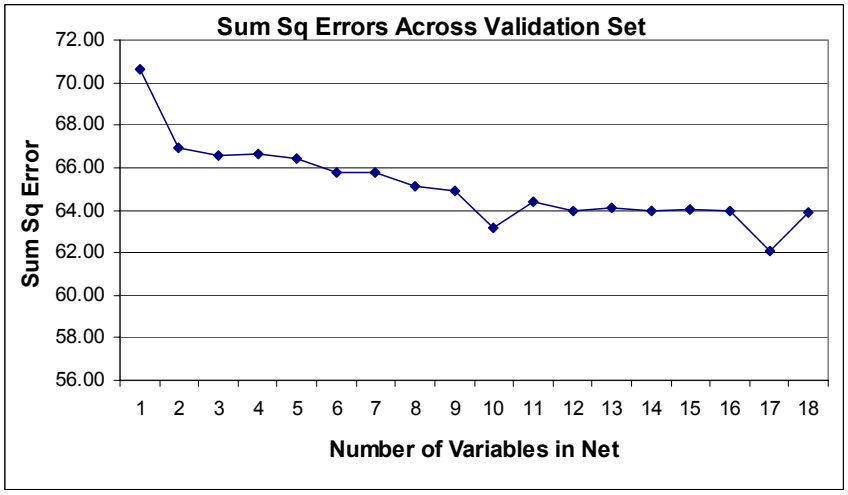


Figure 39. SSE Of Individual Networks

Looking at Figure 39, it appears that SSE is minimized with 10 or 17 features. Additional networks where trained for networks with 9, 10, 11, 16, 17, and 18 features and ROC curves inspected for both validation and TEST sets before the final number of features was selected. Networks with the top 17 features consistently provide the best

results. Although 10 features provide a more parsimonious model, performance across the TEST set is lacking.

Figure 40 presents the performance for a typical 17-feature network across its validation set. It is common practice to end training when a performance metric is optimized on a validation set. This prevents over training the network, which destroys generalizability of the network. Figure 41 presents the same network's performance across the TEST set. It is interesting to note the stark difference in performance between the sets.

The TEST set is expected to yield lower performance as it provides a more realistic picture of how the model could be expected to perform in the population. Performance of the validation set is overly optimistic because the network is optimized for the validation set performance. Given such a drastic decrease in performance across the TEST set, the author is lead to believe the TEST set to be particularly difficult in terms of classification. This realistically leads to the conclusion that performance reported in this research can be expected to be conservative. In any case, these results make a strong argument for the use of truly independent data sets for model validation. Only reporting validation results for models optimized on that set may result in overly optimistic conclusions and possible implementation of an invalid model.

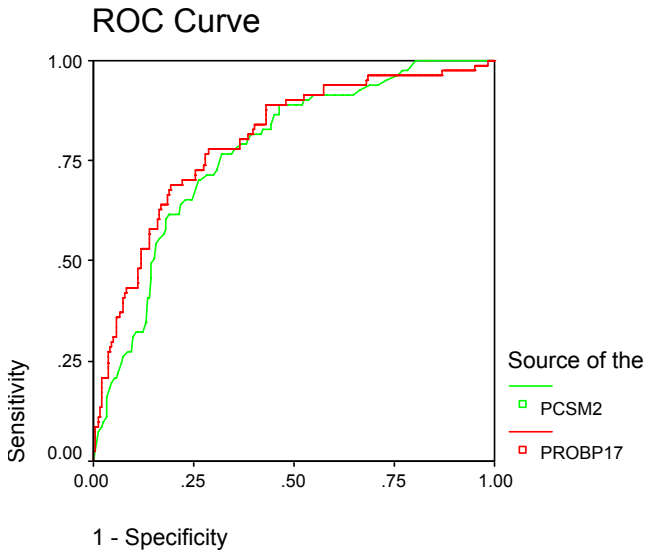


Figure 40. 17 Feature Network Performance Across Validation Set

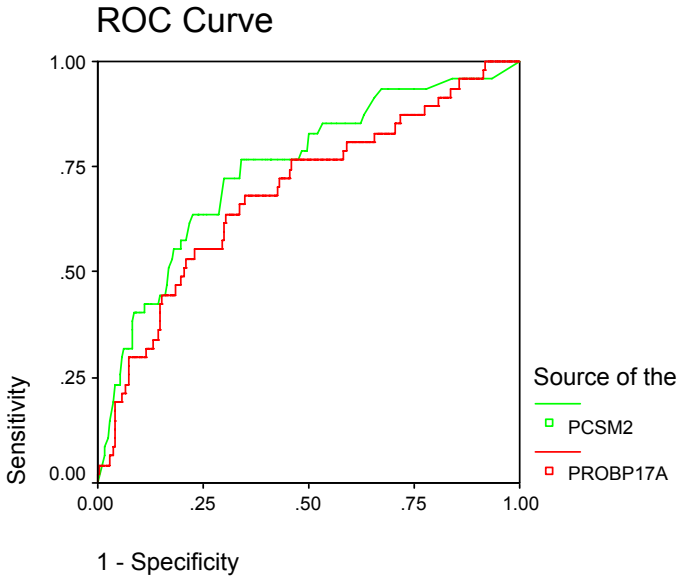


Figure 41. 17 Feature Network Performance Across TEST Set

Once the 17 features were selected, networks were trained on the other two data sets, DATA_B and DATA_C. It was expected that providing the network with a larger

proportion of failures would yield improved results. Such improvements where not found. In fact, performance was degraded.

Figure 42 presents results of a network trained on DATA_B. These results are not valid. Figure 42 is presented to caution the reader duplicating these results. This invalid performance is a consequence of the fact that there is no way within Neural Connections to avoid randomly selecting the repeated bootstrapped sample of failures for use in both the training and validation subsets of DATA_B. Hence, many of the observations in the validation set are also present in the training set. This results in the overly optimistic ROC curve in Figure 42. Figure 43 presents the same network's results for the TEST set. Again, the importance of the TEST set is clearly demonstrated.

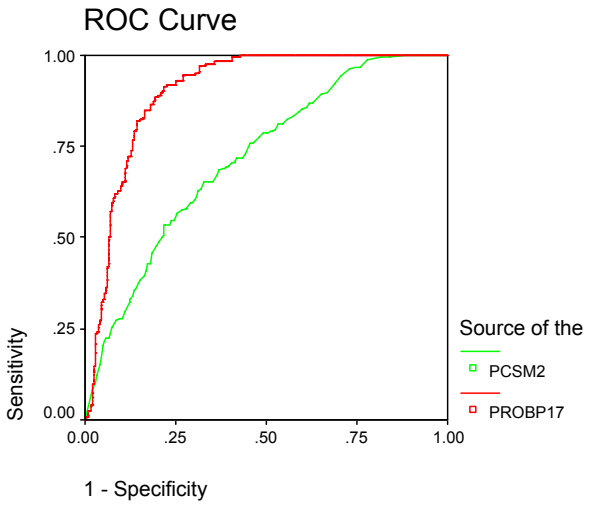


Figure 42. 17 Feature Network Results From DATA_B Across It's Validation Set

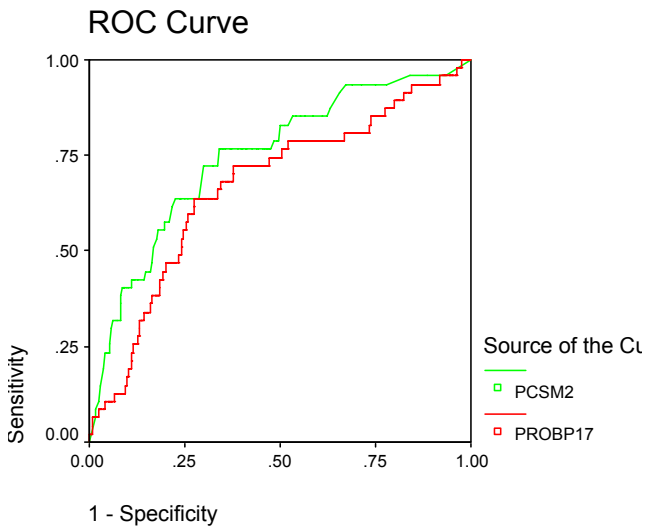


Figure 43. 17 Feature Network Results From DATA_B Across the TEST Set

The final step in optimizing the network is to select the appropriate number of nodes in the network's hidden layer. All networks trained to this point had one node for each feature. Additional nodes provide the network with more classification power. This phenomenon relates to the fact that each node separates the data with a hyper-plane. With enough hyper-planes, the data can be grouped into its separable classes. The reader will recall that given enough nodes, a network can accurately approximate any function. The problem with too many nodes is that the network can begin to approximate a function of the noise within the training data. This leads to a breakdown of generalizability when the network is applied to independent data.

In order to optimize the number of hidden-layer nodes, one network was trained with 4 through 26 nodes, 34 nodes, and 50 nodes. 4 nodes resulted from using Neural Connections capability to automatically find an optimal number of nodes during training. In general, this feature has produced less than desirable results, with the resulting number

of nodes being orders of magnitude less than the number of features. The author prescribes to a general rule of thumb in network architecture in which at least one hidden node is used for each feature.

This iterative network training procedure resulted in further investigation of 15, 17, 19, and 23 node networks. For each of these architectures, a second network was trained. The 23-node architecture appeared to provide optimal, yet stable results. A third 23-node network was trained to further test this stability. Figure 44 presents typical results for a 23-node network with the top 17 ranked features.

It is interesting to note that no single network outperformed PCSM on the TEST set throughout the entire network optimization process. However, very many networks have provided what seem to be optimistic results when validation set performance is considered. A properly designed neural network is considered to be a very powerful modeling tool. The fact that PCSM remains dominant after such a rigorous network architecture design process provides perhaps the strongest validation of the model yet.

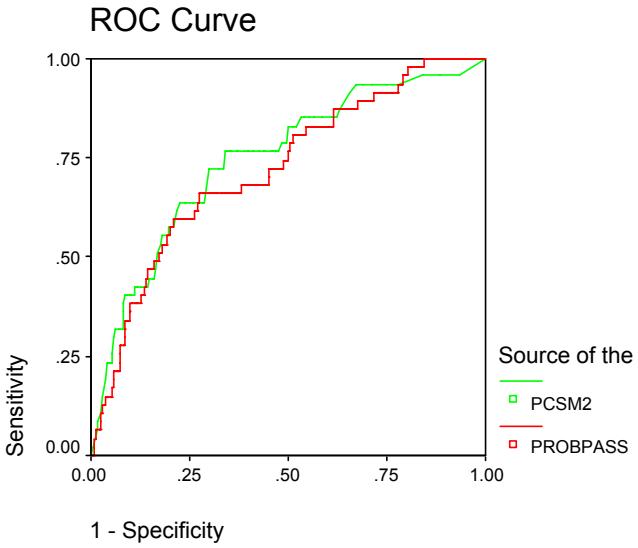


Figure 44. ROC Curve For 17 Features and 23 Hidden Nodes Across The TEST Set

4.4.2 Discriminant Analysis Feature Selection Results

When the assumptions underlying discriminant analysis hold, the discriminant function is highly related to linear probability models in regression analysis (Dillon & Goldstein, 1984). Dillon & Goldstein (1984) point out several philosophical differences in these two types of analysis. Similar to regression analysis, it is necessary to consider higher-order terms for inclusion in the discriminant function in order to obtain the best results. Due to limitations on the number of variables in SPSS, every interaction term could not be considered in one stepwise discriminant analysis run. Therefore, Young (2002) used a heuristic search methodology by which he iteratively investigated every second-order interaction term for the 28 independent variables available in that analysis.

This was accomplished by generating the set of interactions for each of the 28 basic variables. Then 28 stepwise discriminant analyses were performed. Each one included the 28 basic variables and one of the 28 sets of interactions. At each iteration,

all significant interactions were retained for a final stepwise discriminant analysis, which was performed on the 28 basic variables and the retained interactions from the 29 previous iterations. This heuristic involved 29 separate stepwise discriminant analyses.

Given the fact that the data used by Young (2002) is a sub-set of the data used in this research, a discriminant function is derived using a set of variables similar to those in Young's (2002) final discriminant function. Due to differences in variable coding decisions in this research and the exclusion of the variable identifying sex, it is not possible to recreate every variable exactly. The discriminant function with the Young-like (2002) variables performed as well or better than the optimized network. Table 27 presents the 22 variables used in this discriminant function. Figure 45 presents a ROC curve of this discriminant function applied to the DATA_A TRAIN set. Figure 46 presents the same discriminant function applied to the TEST set.

Table 27. Discriminant Function Variables Based on Young (2002)

1.Pilot x FixSgle	7.Pilot	13.ANGAFR x PS2X2	19.Pilot x ANGAFR
2.BAT_Age x ROTC	8.ROTC	14.Quant	20.Quant x ANGAFR
3.AERO1	9.H2CX1	15.PS2X2	21.ITMR
4.AERO2	10.ANGAFR x H2CX1	16.ANGAFR	22.BAT
5.AERO34	11.MultiInstrum	17.BAT_Age	
6.Pilot x Pilot	12.GPA	18.ROTC x MultiInstrum	

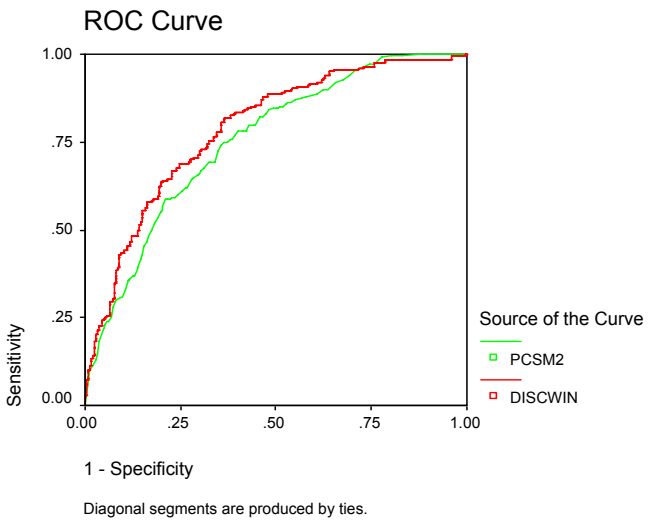


Figure 45. DATA_A Discriminant With Young (2002) Variables

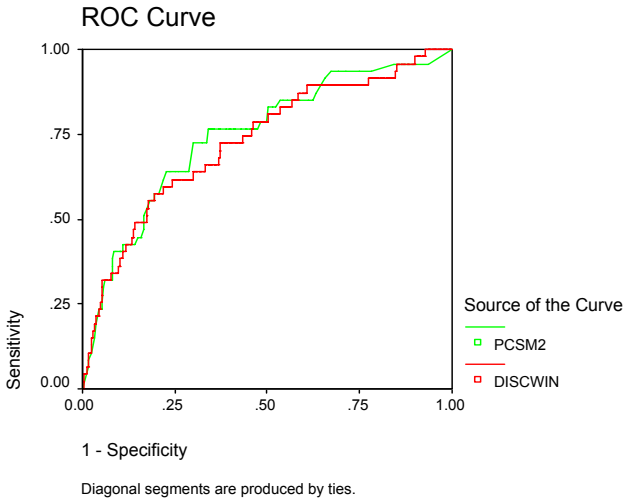


Figure 46. DATA_A Discriminant On TEST With Young-like (2002) Variables

In an effort to reduce the number of interaction terms that require investigation, a new approach is also taken. A reduced set of basic variables is pre-selected in order to minimize the number of interactions. To do this, the list of ranked salient features in Table 24 is compared to the list of variables significant in a stepwise linear regression on

the DATA_A set. These regression variables in Table 28 are presented in the order they entered the regression. This regression was performed on the combined DATA_A and TEST sets. No variables left the regression once they entered. Judgment was used to select a parsimonious set of 14 basic variables. The variables are presented in Table 29. Note some differences in variable names compared to Table 15 are due to changes in coding of dummy variables for independent model development. Such changes simply involve the combination of multiple levels of nominal or ordinal variables into fewer dummy variables. New variable names relate to the variables combined.

All 91 interaction terms were created for these 14 variables and added to the set of basic variables. SPSS was able to perform stepwise discriminant analysis on the entire set of 105 variables (basic variable and interactions) with probability of F-ratio to enter/leave set to 0.05/0.10. The resulting discriminant function included 20 variables. The 20 variables are listed in Table 30. The numbering of these variables in Table 29 does not indicate a ranking or significance among the variables. The numbering of the variables relates to the naming of the interaction variable used during the creation of interaction terms. By comparison, Young's (2002) discriminant function included 25 variables after investigating 378 interaction terms for 28 basic variables.

Table 28. Regression Variables For Selecting Basic Discriminant Variables

DATA_A TRAIN Significant Regression Variables: 0.15 to enter, 0.20 to leave				
1. Pilot	4. H2CX1	7. TMSAV	10. BAT	13. AIAP
2. PS2Z2	5. ROTC	8. ANGAFR	11. AIAR	
3. AERO2	6. BAT_Age	9. ITMR	12. PS2Y2	

Table 29. Basic Variables For New Discriminant Analysis

Name	Number
PILOT	1
AERO1	2
AERO2	3
MULINSTM	4
BAT	5
H2CX1	6
PS2X2	7
PS2Y2	8
PS2Z2	9
TMSD	10
ITMR	11
BAT_AGE	12
ROTC	13
ANGAFR	14

Table 30. New Discriminant Function Variables

1. Pilot	11. MULTINSM x BAT AGE
2. AERO2	12. BAT x BAT AGE
3. H2CX1	13. H2CX1 x H2CX1
4. BAT AGE	14. H2CX1 x PS2Z2
5. Pilot x AERO2	15. H2CX1 x ITMR
6. Pilot x ITMR	16. H2CX1 x BAT AGE
7. Pilot x ROTC	17. PS2X2 x PS2Z2
8. AERO1 x PS2X2	18. PS2Y2 x TMSD
9. AERO1 x ITMR	19. TMSD x ROTC
10. AERO2 x ITMR	20. ROTC x ROTC

Figure 47 presents the discriminant function derived on the DATA_A set applied to it. Figure 48 presents the performance of this discriminant function applied to the TEST set. Note the stark contrast in performance when the independent data set is used. It appears that the discriminant function based on the variables in Young's (2002) final

discriminant model provide much better performance. However, neither discriminant function is capable of out performing PCSM on the independent TEST set.

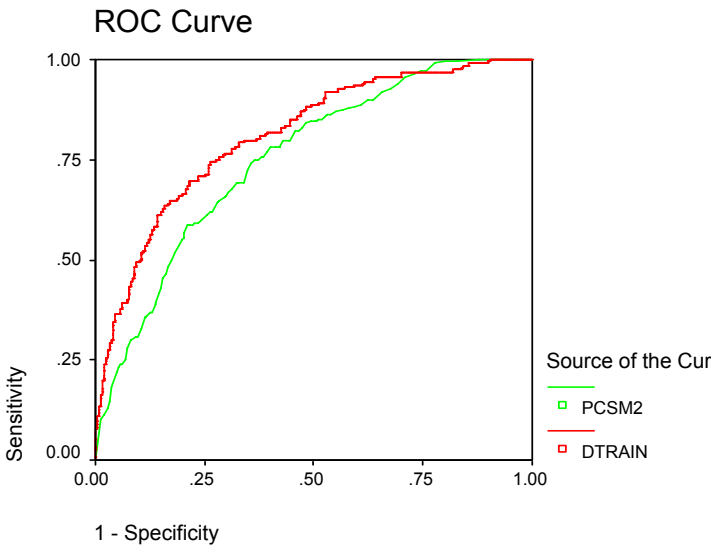


Figure 47. New Discriminant Function Applied To DATA_A TRAIN

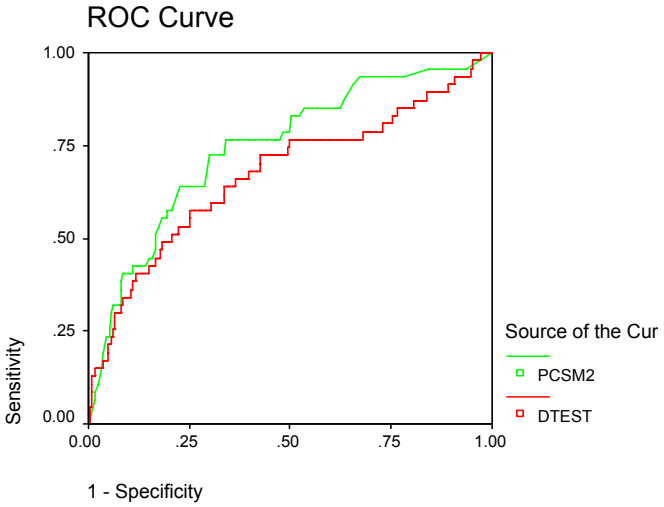


Figure 48. New Discriminant Function Applied To TEST

4.4.3 Ensemble Method Results

Perrone & Cooper's (1992) Ensemble method provides a way to combine model outputs in an attempt to improve performance. The Ensemble method utilizes model error correlations for different models to derive the weights of a linear combination of model outputs. The Ensemble method can take advantage of multiple local minima on the error surface; hence it works best when combining dissimilar models. One method of generating such dissimilar models is to train models with drastically different sets of initial weights. The most dissimilar models can be identified using factor analysis of a matrix of the different model's errors (Young et al., 2003). This procedure will identify the models explaining the latent constructs underlying the orthogonal factors of a varimax rotation. Highly similar models will have high factor loadings within a common factor, hence only one feature is selected from each factor.

Three (3) networks of each of the 3 activation functions available in Neural Connections where trained on the DATA_A TRAIN set. The random number seeds controlling the training/validation set assignments and the initial weight distribution was changed for each of the 9 networks. Each network was applied to the TEST set and a matrix of TEST set errors was constructed.

Factor analysis was performed on TEST set matrix of errors. Ideally, the eigenvalues of several factors will be greater than 1.0. However, in this analysis the first factor accounted for 97.6% of the variance. Hence, the Ensemble method was unable to glean additional performance by combining networks that represented factors accounting for an insignificant proportion of variance. The result was that the networks selected to represent the second and third factor had substantial factor loadings on the first factor.

Thus, these two networks where highly similar to the network representing the first factor. The result was that all entries of the correlation matrix of these errors are near unity. This limited the contribution of the Ensemble method severely by reducing it to a matter of averaging the outputs of nearly identical models.

Table 31 and 32 present the unrotated and varimax rotated factor loadings across the first 3 factors of the matrix of errors from the 9 networks, respectively. Note that in the unrotated factor loadings, all 9 networks load heavily on the first factor. The varimax rotated factor loadings provide no additional insight because each network is essentially providing the same model. Table 33 provides the correlation matrix for the errors matrix of the three networks combined with the Ensemble method. In Table 33, the "E" denotes the use of errors rather than actual model outputs. SIG refers to networks trained with a sigmoidal activation function. Likewise, HYP and LIN denote hyperbolic-tangent and linear activation functions, respectively.

Table 31. Unrotated Factor Loadings of Errors Matrix

	Component		
	1	2	3
ESIG1	.984	-.035	-.123
ESIG2	.989	.094	-.058
ESIG3	.991	-.066	.021
EHYP1	.990	.052	-.022
EHYP2	.992	-.039	-.030
EHYP3	.976	.138	.148
ELIN1	.989	-.106	.064
ELIN2	.990	.057	-.073
ELIN3	.988	-.093	.075

Table 32. Varimax Rotated Factor Loadings Of Errors Matrix

	Component		
	1	2	3
ESIG1	.677	.575	.443
ESIG2	.650	.500	.563
ESIG3	.563	.644	.505
EHYP1	.615	.543	.557
EHYP2	.608	.610	.495
EHYP3	.488	.521	.697
ELIN1	.521	.686	.501
ELIN2	.656	.524	.532
ELIN3	.514	.679	.515

Table 33. Correlations For Errors Matrix

	ESIG1	**EHYP3**	**ELIN1**
ESIG1	1		
EHYP3	.95	1	
ELIN1	.97	.96	1

Figure 49 displays the result of combining the three networks with the highest varimax rotated factor loadings on each of the three factors in Table 32. The combined model is compared to two of its three component networks. It can be seen that in this case the Ensemble method performs worse than the selected linear network. Similar issues caused inadequate performance when combining the outputs of the discriminant functions with one or more network models. In this case, maintaining multiple models for the purpose of employing the Ensemble method would be cumbersome at best with

little if any performance increase. The simplest approach is to select the single best model as the final model.

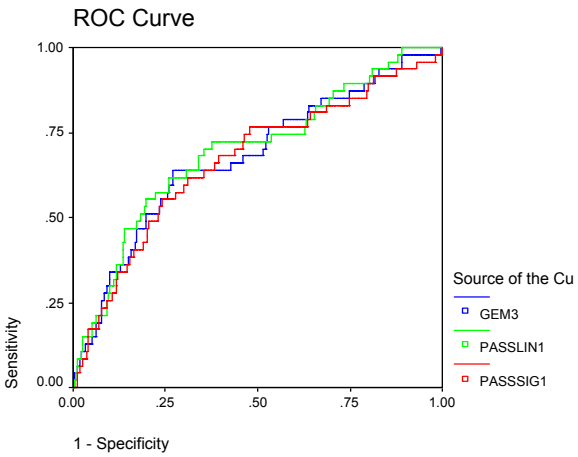


Figure 49. Ensemble Method Results vs Individual Networks

The discriminant function based on the Young-like (2002) final discriminant model is the best model overall. It is interesting to note that the discriminant function outperforms the optimized MLP neural network in this application. Despite the rather involved process required to produce the 7 PCSM inputs, PCSM is much more parsimonious than this discriminant function with 22 variables. A key component of the discriminant function that makes it a possible surrogate for PCSM is the fact that it involves a single vector of discriminant weights. The optimized network on the other hand involves the application of both a 17 x 23 matrix of inputs weights (not including a bias term) going into the hidden-layer, and a 23 x 1 vector of weights after the hidden-layer. Without significant performance gains, such intricacies make implementation of a neural network unrealistic for AETC in terms of presenting and explaining the model to the end user.

4.5 Model Comparison Results

As a model nears perfect classification, that is classifying all observations with zero error, the model output would move from the continuous to binary domain. All failures would receive an output of zero and all passes an output of one. A very good model would then be expected to assign all failures an output near zero and all passes an output near one. Figure 50 provides a notional picture of how the cumulative distribution of failure and pass outcomes might be assigned for the ideal model.

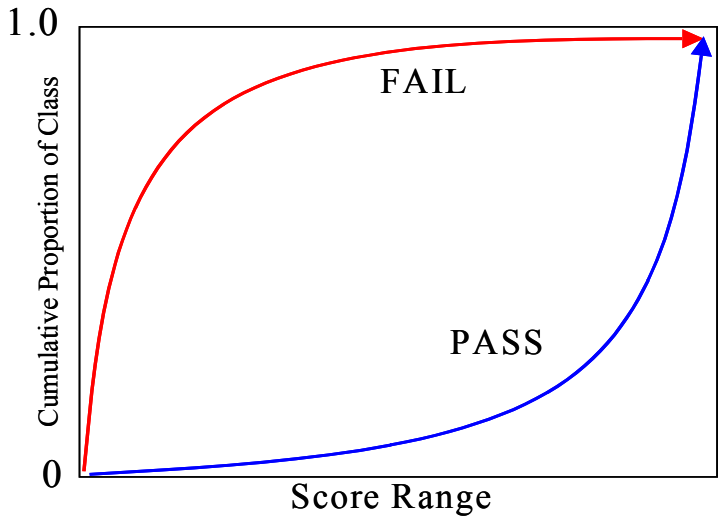


Figure 50. Cumulative Proportion of Classes for the Ideal Model

When applied to an independent TEST set, a ROC curve provides an estimate of the model's generalized performance across the entire range of classification thresholds for the two most common metrics of interest, probability of target detection and probability of false alarm. Although ROC curves have been used as the primary means

of displaying and comparing the different model's performance to PCSM, the next several charts attempt to see how the logistic regression and the 22 variable Young-like (2002) discriminant model compare to PCSM in terms of emulating this notion of an ideal model in Figure 50.

The following results relate to the TEST set. Figures 51 and 52 present how the logistic regression model outputs are distributed across the TEST set passes. The logistic regression is chosen over the linear regression because of the expectation that the sigmoidal activation function will aid in separating the distributions of passes and failures across the score range. Figure 51 provides actual counts in each score range, while Figure 52 provides the same information as a cumulative proportion of the sample of passes for comparison to our notion of an ideal model. Figures 53 and 54 provide the same information for the sample of failures in the TEST set. Figures 55 through 58 present the same information as Figures 51 through 54 for the 22-variable Young-like (2002) discriminant function. With respect to passes and failures, the PCSM results are unchanged in this series of figures.

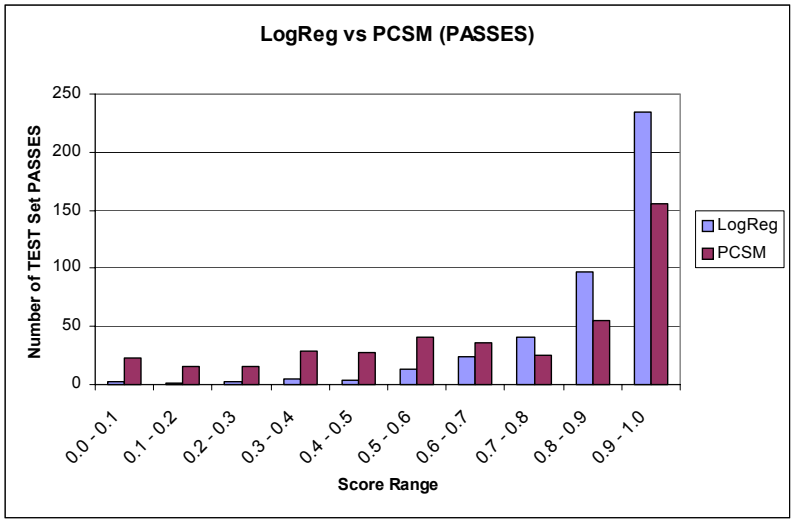


Figure 51. LOGREG TEST Set Passes Distribution

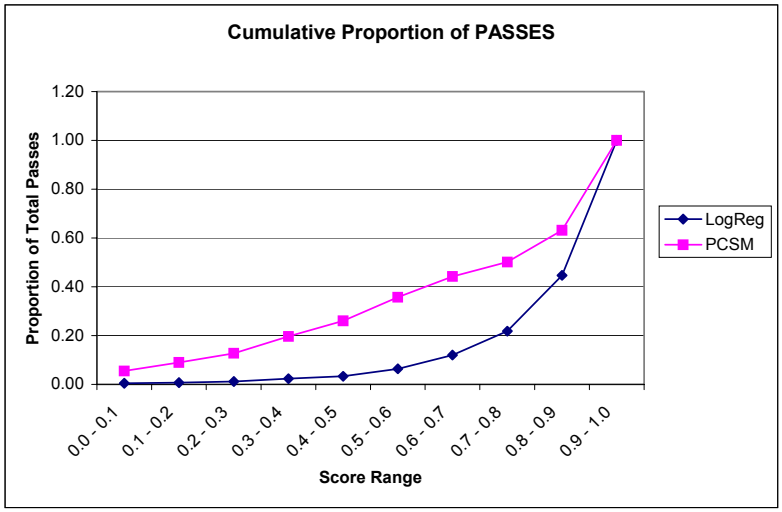


Figure 52. LOGREG Cumulative Proportion of Passes

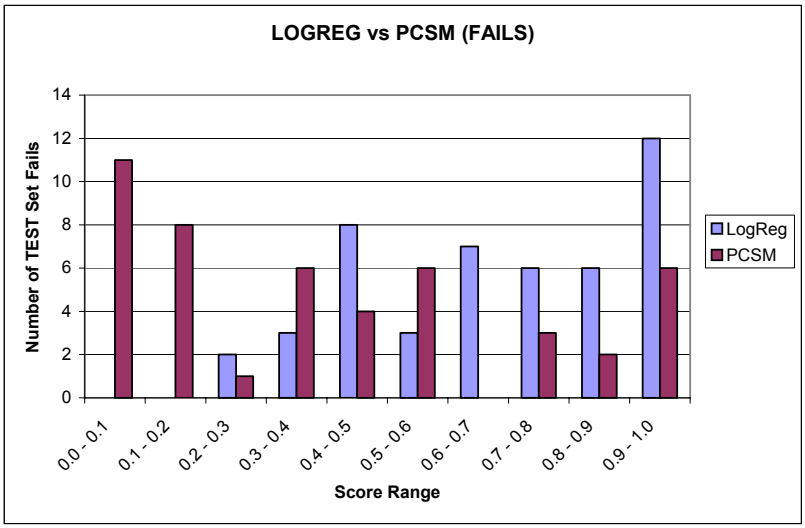


Figure 53. LOGREG TEST Set Fails Distribution

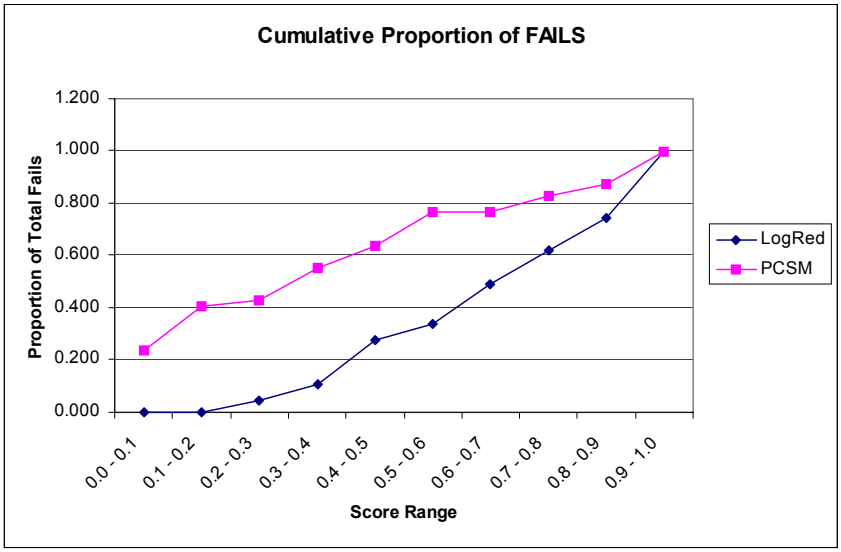


Figure 54. LOGREG Cumulative Proportion of Fails

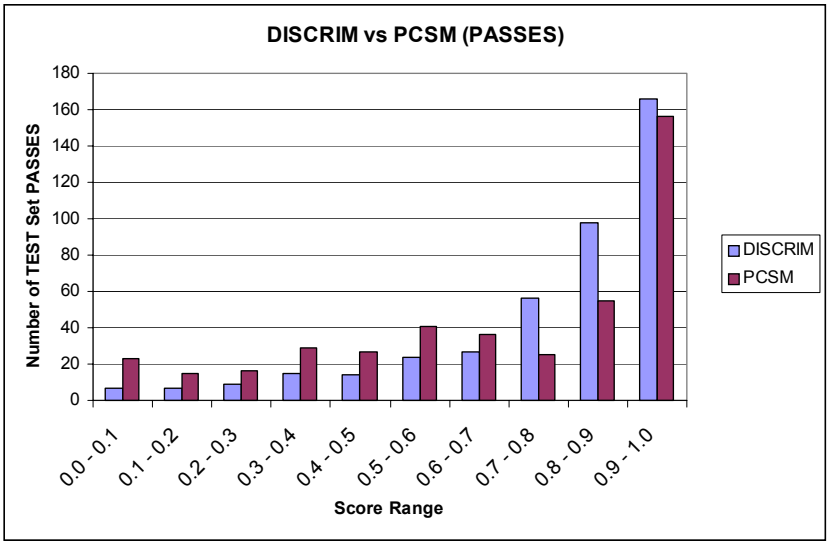


Figure 55. DISCRIM TEST Set Passes Distribution

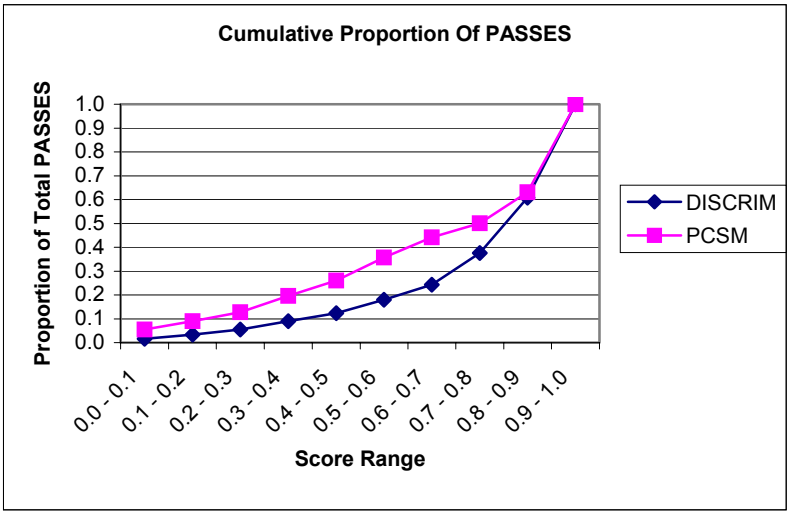


Figure 56. DISCRIM Cumulative Proportion of Passes

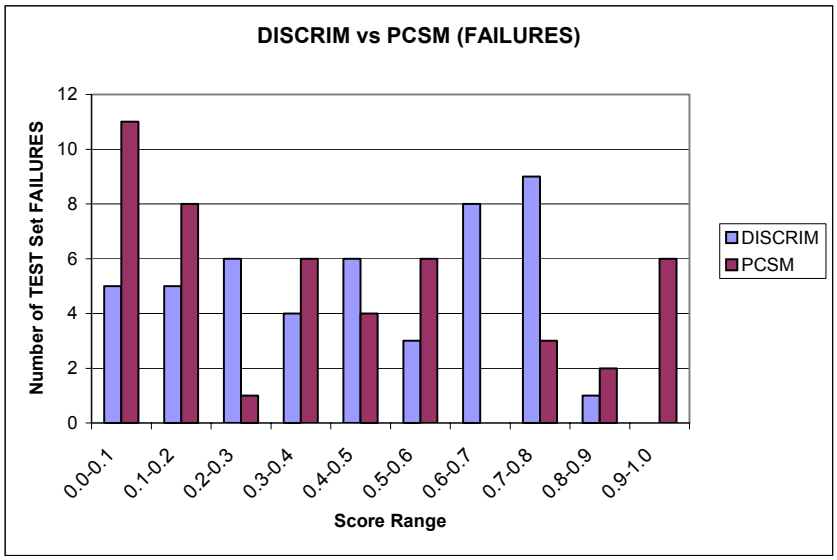


Figure 57. DISCRIM TEST Set Failures Distribution

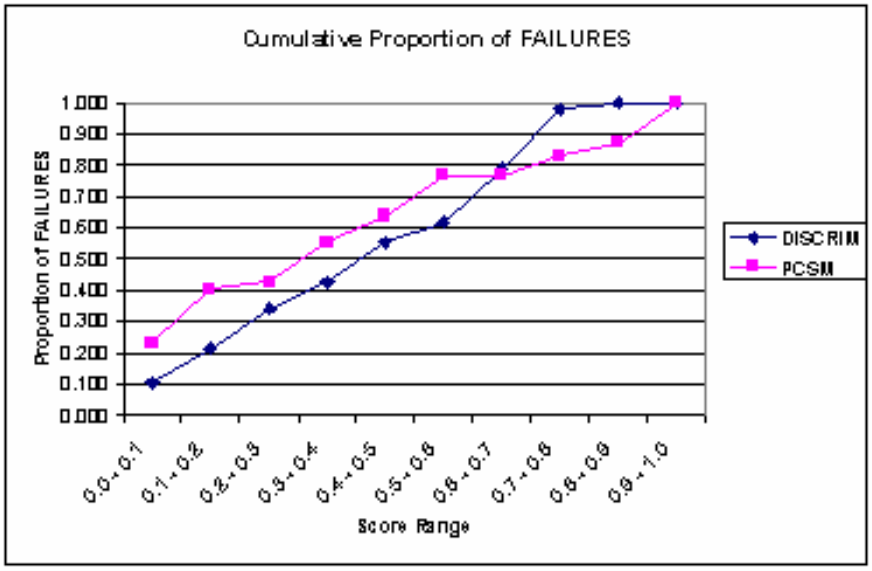


Figure 58. DISCRIM Cumulative Proportion of FAILURES

For the passes, both the logistic regression and discriminant model are more like the ideal than PCSM. In terms of the failure sample, PCSM yields better performance in both cases. However, neither model adequately models the ideal for the failures. Herein lies the crux of the problem as it relates to predicting UPT failure. The distribution of predictor scores for the failures is practically uniform. This indicates that reasons for failure may not be related to constructs measured by PCSM. Hence, predicting failure is a more difficult problem for reasons that may not be related to measures of ability; at least not for the measures of ability, which explain the constructs underlying the current data set. If a predictor with construct validity for a construct that accounts for failure, it would certainly show up in the factor analysis.

4.6 Chapter Summary

This chapter presented the results of this research as they relate to achieving each of the 3 research objectives. All three parts of this research validated the current PCSM model. The work in the area of updating PCSM's regression weights showed that the updated weights are no better than the current weights. Linear and logistic regressions performed almost identically, regardless of whether the undefined discrete sigmoidal function approximation is applied to the outputs. Updating the EQPMOT standardization resulted in performance degradation. PCSM also performed well against independent models developed using powerful multivariate techniques such as discriminant analysis and neural networks. Differences in the cumulative proportion of passes and failures across the PCSM score range may provide some insight into how PCSM performs differently for passes and failures.

V. Conclusions

5.1 Introduction

This chapter discusses conclusions drawn from each aspect of this research. In order to help future researchers, lessons learned through this effort are presented. New insights gained through working with the data and PCSM model are discussed. Finally, recommendations for future research are suggested for consideration by future researchers.

5.2 Literature Review Findings

The pitfall receiving the most attention in this research is the effect of range restriction on predictor/criterion correlations. A wealth of research is available on the range restriction problem. The literature review in this research brought together much research aimed at understanding the accuracy of range restriction corrections. Most concentrated on accuracy when the linearity and homoscedasticity assumptions are violated. Studies involving both empirical data and Monte Carlo simulations are reviewed.

Study results show that, in general, the corrections tend to be negatively biased. In fact, Bobko (1983) showed this expectation theoretically. Accuracy is largely dependent on the severity of distributional assumption violations and selection ratio. Except for the most severe violations, applying Lawley's multivariate correction would be expected to produce benign results. On the other hand, no study reviewed in this research specifically argued that failing to perform the correction routinely causes

adverse results. The implication is that range restriction may play some, yet unquantified, role in selection of a non-optimal set of predictors. No real world examples were found to suggest that variables selected based on corrected correlations would differ from those selected using range restricted correlations. When the magnitude of the uncorrected correlation is large, few examples of corrected correlations with near zero bias were found. Most instances occurred at a specific experimental point and not across a range of tested points.

Estimates of the unrestricted population mean, standard deviation, and predictor inter-correlations are required to perform the Lawley correction. Such estimates are often not available or their estimates are of questionable accuracy. No study reviewed investigated the accuracy of corrections when unrestricted population estimates are varied across a significant range. Furthermore, no study looked at which population estimates drive accuracy for PCSM related data. Hence, any correction based on such estimates is suspect. From my research, I believe range restriction affects all potential predictors for the PCSM model to some degree. Therefore, a model based on uncorrected correlations is likely to result in a model that includes an optimal set of predictors. These issues leave corrections open to possible criticism.

Although the corrections have been used in several PCSM studies, no specific argument for using the correction for the PCSM data was found. The predictor/criterion correlations in the PCSM model are certainly range restricted; however, no case is made as to how well the PCSM data meets the underlying assumptions or how the extreme selection ratio that exists would tend to affect correction accuracy. No attempt known to the author has been made to estimate the accuracy of corrections on PCSM data.

Differences in the performance of PCSM related models based on corrected correlations should be investigated.

5.3 Methodological Conclusions

The PCSM database does not have adequate configuration management. A mechanism is needed to ensure the integrity of the PCSM database. Many cases of inconsistent data exist. For example, there exist records of pilots certified at some level (pilot, instructor, commercial) who have zero reported flying hours. The database entry system should have safeguards that prevent entry of inconsistent records. Apparently, no document accompanies the PCSM database in which the data available, calculations, transformations, etc. are concisely explained. A standardized and unique identifier must be put into place for all data tables. Efficient and accurate queries of the multiple databases depend on it. There exist multiple UPT performance databases or versions of them for each year. Further, many tables exist within each database with no available explanation as to what each represents.

A wholly independent TEST data set must be used to properly verify model performance. The Neural Connections software used in this research provides the capability to select training, validation, and test sets from the data used in model development. As the model trains, its performance is optimized for the validation set. Hence, the model's generalizability is questionable. To report only validation set results is naive at best. An even more problematic situation would occur if the researcher combined results from the training and validation sets together. This would be expected to give an even more optimistic result than would validation results when a model is iteratively trained by optimizing performance on that validation set.

Finally, if several different data sets are being used to train independent models for the same purpose, each model should be compared via the same wholly independent TEST set. Such is the case in this research. These dangers have been made abundantly clear through the presentation of results for both validation sets and the unique TEST set.

Future researchers may find it interesting to note that PCSM performed differently on the TEST set in this research and the entire data set used in the work of Young (2002). By locating a specific combination of the probabilities of detection and false alarm for PCSM on the ROC curves in Figures 59 and 60, it is apparent that PCSM performs better on the TEST set in Figure 60. In Figure 59, the lower line presents PCSM performance on the combined training and validation set used by Young (2002). It appears that PCSM performed better in this research, which marks the need to estimate PCSM's true generalized performance.

Given that the Young (2002) data is thought to be a subset of the current data, one would not expect such results. This comparison of PCSM results across different data sets is believed to be valid because the PCSM model is unchanged in these two applications. However, this research included the addition of failures for academic and military reasons whereas Young's (2002) did not. Total failures where about 10% in both data sets. One should also keep in mind that the TEST set included 470 records, while Young's data included over 1,700 records. If PCSM performance is sample dependent, then one must judge the performance of both PCSM and an independent model with statistical procedures, which account for error. For example, Young (2002) calculated a 95% confidence interval at several locations on his ROC curve to show a statistical difference in model performance. The disparity in proportions of failures and

passes necessitates caution when considering the Young's (2002) confidence intervals. Confidence intervals generated for the probability of target detection for a fixed probability of false alarm would be larger due to the much smaller failure proportion in the data.

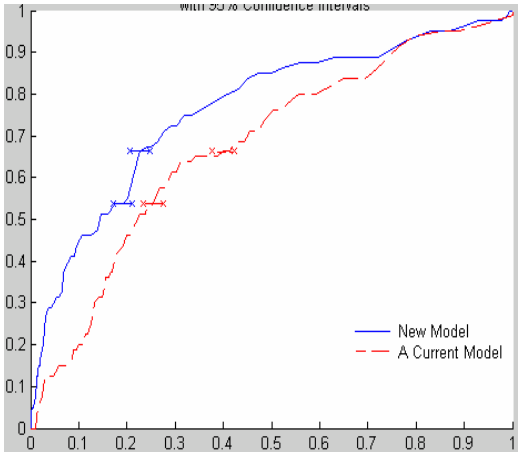


Figure 59. PCSM Performance Across Young's (2002) Data

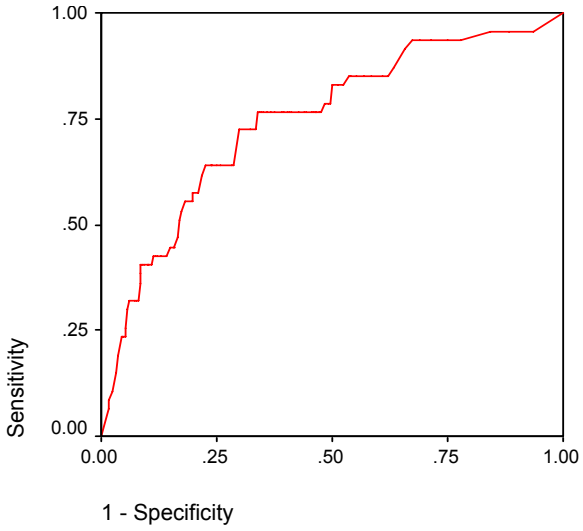


Figure 60. PCSM Performance Across the TEST Set

Young's (2002) results in Figure 59 are not without other methodological issues. Apparently, Young did not use an independent TEST set. Since it is known that Young (2002) used Neural Connections capability to randomly select training and validations sets for each network, any results for a validation set is considered to be optimistic in terms of generalization. This is because Neural Connections optimizes the trained network for the validation set. Furthermore, Young's (2002) results combine model outputs for both training and validation sets. This further optimizes the model performance and degrades confidence in the model's generalized performance.

The methodology in which Young (2002) performed factor analysis to select models for use in Perrone & Cooper's (1992) Ensemble method also raises concerns. All model errors in the factor analysis must represent each model's output for the same record. From Young's (2002) thesis, it is believed the 10 models trained for this purpose used a different random number seed, which assigns the training (70%) and validation (30%) sets. Social security number was then used to match the 10 model outputs with the correct target value for each record to facilitate model error calculations. Hence, for a specific record, approximately 7 of the 10 errors recorded reflect performance for the training set. Once the three models where chosen to apply the Ensemble method, 70% of the outputs within each model represent training set outputs. In short, this produces a result that estimates performance on a generalized training set rather than performance in the population.

In summary it is believed that Young's (2002) results are exceedingly optimistic for two reasons. First, due to the chance that that PCSM's true generalized performance is better than seen by Young (2002). Second, due to the arguments suggesting Young's

(2002) model's performance is overly optimistic. If both issues where addressed, the performance of the two models could move toward each other. Such expectations make PCSM's dominant performance in this research more believable. However, such expectations may not hold and thus are not an estimate of how good Young's (2002) model really performs. After all, the best model in the present research is based directly on the final discriminant function in Young (2002). Young's model should be validated on a truly independent test set.

5.4 Validation Study

The current PCSM variables have been shown to have construct validity in that they consistently represented the most significant factor analysis factors. Factor interpretation was unhampered by implementing orthogonal varimax rotation. It fact, the same interpretations that would be made with unrotated factor loadings become much more apparent after varimax rotation. In this research, rotation is believed to have strengthened the validation study. Factor interpretations are upheld by the demonstrated results over a variety of methods such as partial correlations, correcting correlations for range restriction, and two types of regression. PCSM performance is dominant in all ROC's based on the TEST set.

5.5 Regression Update Conclusions

Two new regressions are derived across multiple data sets. Each displayed similar performance to PCSM, yet none outperformed PCSM on the TEST set. Furthermore, a true logistic regression did not outperform the linear regression. The requirement for applying the discrete sigmoid approximation to the results of a linear

regression in the current PCSM model was not shown to be beneficial. The origin of this discrete sigmoid approximation is unknown. AETC should reevaluate this aspect of the PCSM model. Perhaps a better alternative would be to simply rescale the linear regression outputs to the desire range of 0.0 to 1.0.

The work performed to gain insight into the current PCSM model shows that the Pilot, FltHrCd, and EQPMOT inputs drive performance in this application. This provides evidence that a more parsimonious model may be available. AETC should investigate the necessity of the four other PCSM inputs. Furthermore, the validity of the non-EQPMOT BAT scores should be reevaluated. Perhaps such an investigation would lead to changes in the BAT test, which capitalized on the predictive value of the EQPMOT input.

Future research in the area of PCSM should begin with a review of the history of the model's development and configuration management. Many questions about the current configuration of the PCSM model remain. The author is unaware of the existence of a set of documents that preserve and explain the history and current state of PCSM. Continued use of PCSM makes such a study worthwhile.

5.6 Independent Model

Despite the complexity & many processing steps involved in the current PCSM model, it still outperforms the independent models when applied to an independent TEST set. For the current data set, discriminant analysis provided as good or better performance than the more complex neural network model. The discriminant analysis feature selection algorithm was shortened relative to Young (2002). A more rigorous

feature selection methodology would likely improve the results of the discriminant function. As is the case with multiple linear regression, the coding of variables can have significant impacts on model performance. Although, the variable coding decisions seem logical and where made after significant inspection of the data was performed, optimization of variable coding was not specifically addressed. Smith (1996) discusses alternative coding schemes for neural network applications. Perhaps investigation into alternate coding schemes would result in better performance.

Using SNR ranked variables to help pick the 14 basic variables in the discriminant model may have been a poor choice. The discriminant function is a linear model. Presenting the model with variables known to be significant in a non-linear model such as a neural network may have introduced noise in terms of separablility. However, there does exist much similarity in the variables found to be significant in the linear regression and by the SNR method.

5.7 Relevance of Research

The applicant population is quite diverse in terms of education and ability levels. The selection process is purposefully and rightfully safeguarded against discriminatory selection policies. The current selection process favors the "whole person" concept of selection over strict measures of ability. This results in a selected sample whose predictor score distributions covering a large portion of score range for most predictors available. Such distributions of ability measures across failures and passes could be a significant contributor to the complexity in prediction of UPT failure. In fact, tests on the difference in predictor score means between failures and passes fail to reject the null for

most predictors available in this data. This result holds when smaller random samples of the data are selected in order to limit the power of the test.

Inspection of Figures 53 and 57 in Section 4.5 reveals that the TEST set sample of failures (N = 47) across the PCSM score range is quite uniformly distributed. Figure 61 presents the distribution of PCSM scores for the 18,927 valid records covering a period of 1993-2001 in the PCSM database. Figure 62 presents the proportion of UPT selections across the PCSM score range for the same 18, 927 PCSM scores.

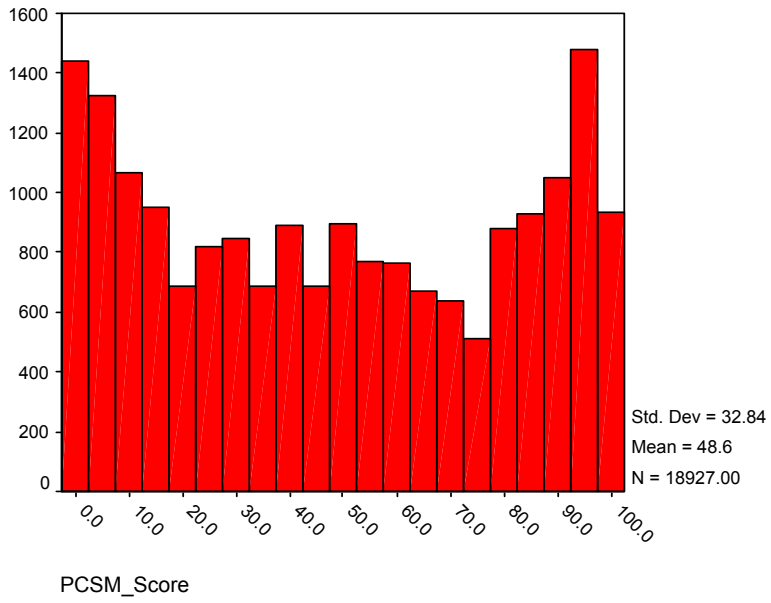


Figure 61. Histogram of PCSM Scores Among Those Selected for UPT

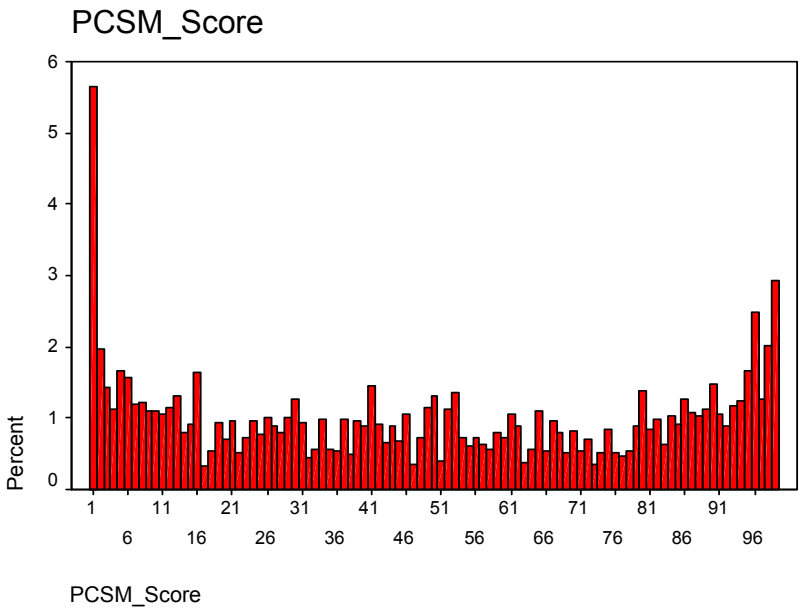


Figure 62. Proportion of UPT Selections Across PCSM Score Range

One need not be privy to Weeks (1998) study on the selection policies of the top four UPT selection sources to draw the conclusion that PCSM is not a significant factor in the selection process. The fact that the distribution of PCSM scores among those selected is not at all what one would expect for a valid prediction model should not reflect poorly on PCSM. Rather, the attrition rates currently experienced by the Air Force may be a result of the selection process. Until more definitive predictive tests are developed, PCSM should be given serious and diligent consideration by every UPT selection board. However, no standardized test should be the sole criterion in such important matters. The officers selected for UPT today represent to a large extent the leaders of tomorrow's Air Force. Leadership and Officership entail much more than the skills required to be a pilot. This aspect is vital to the UPT selection process; it cannot and should not be totally removed.

The fact that scientifically applying very powerful statistical techniques in this analysis with little to no improvement on the current PCSM model is strong validation for PCSM. Although its development and validity are somewhat esoteric, even to those associated with the program, the model is a valid predictor of UPT success. AETC/SAS has long encouraged UPT selection boards to make significant use of the information provided by the PCSM score. Unfortunately, this is not the case for the AFA and ROTC, who select a majority proportion of UPT candidates. AETC should consider enforcing a combination of minimum PCSM score standards and the "whole-person" concept of selection for all applicants.

5.8 Recommendations for Future Research

A possibility that should be considered is to remove the flying experience input from the PCSM model and break the selection process up into three stages. At the first stage, a minimum PCSM score qualification standard would be added to the AFOQT minimum score standards already in place. The second stage would involve applying a standardized selection process or the independent selection processes currently in place. With low PCSM scores removed, the selection boards would still be free to use the PCSM score as they see fit within the "whole person" context. The third and final stage would involve eliminating candidates based on performance in the 50 hours of pre-UPT flight training currently provided by the Air Force to those not possessing a private pilot's license.

It should be noted that the AFA is represented by only 56 of the 3,343 records used in this research. Therefore, no meaningful conclusions can be made as to the

performance of the discriminant model or PCSM when applied to the AFA applicant population. A separate study should be conducted to investigate the performance of PCSM in the AFA population.

Currently, the AFOQT attempts to measure flight related skills in a pencil and paper format. The results of such tests undoubtedly have strong correlation with an applicants flying experience. This should be eliminated and properly placed in the context of a test similar to the BAT. Such a test would improve upon the reliability of measuring the ability to learn complex flying tasks over a period of time under differing workloads. The test would also incorporate the more academic skills required for flight that are suggested for removal from the AFOQT. The test could provide basic instruction in the skills required, followed by practice and correction in an environment that is more realistic than the current BAT test, yet simple enough to make it cost effective. To this end, the latest version of the BAT test incorporates pedals for simulating rudder controls and is expected to become operational soon (Pugh, 2003).

Weeks (1998) theorized that attrition is linked to the availability of instructional resources, performance feedback, and additional time to hone the skills required in flight. Hence, the problem of success at UPT may not be so much about a specific flying ability or even some esoteric notion of intelligence, but rather the pace at which one learns the required skills in the presence of competition for the limited instructional resources at UPT. An objective of an improved BAT test, such as the one imagined here, would be to measure how an applicant performs given the dynamic theorized by Weeks (1998). If successful, such a test would certainly provide the promise of accurately predicting UPT outcomes.

The fact that PCSM's validity remains intact when compared to very powerful classifiers such as neural networks and sound statistical techniques, begs for further investigation into the model itself before more independent models are developed. Along this vein, a design of experiments approach coupled with regression analysis could provide valuable insight into the PCSM model. This approach could investigate the necessity of the 4 PCSM inputs that appear too extraneous in terms of PCSM performance on the TEST set in this research.

It seems evident that those who fail UPT, do so for reasons not measured by PCSM or known to the selection boards. The fact that PCSM scores for those who fail UPT are uniformly distributed across the score range indicates that some unknown trait related to failure occurs without regard to one's PCSM score. One might investigate the variables that represent latent constructs underlying the passing and failing groups separately. This could be accomplished by performing factor analysis on each group. If factorial invariance does not hold, insight may be gained into the differences between those who pass and those who fail UPT.

Searching for a psychological reason or factor related to failure may provide insight into the development of valid predictors of failure. Specialized interviews could be developed and administered before and after UPT training. Failures could be administered an interview which specifically focuses on identifying the reasons for failure.

A final avenue of suggested research could involve an attempt to estimate PCSM's generalized performance presented in ROC format. Such research is prompted by the fact that PCSM performance is to some degree dependent on the proportion of

failures present in a sample. The estimate could be accomplished by collecting a large number of PCSM data samples with differing failure proportions via bootstrapping. PCSM's performance in terms of the ROC parameters could be sampled across the score range for each bootstrapped data set. Confidence Intervals could then be built around a discrete set of points along the mean ROC.

Appendix A. Matlab Code for Bootstrap Resampling

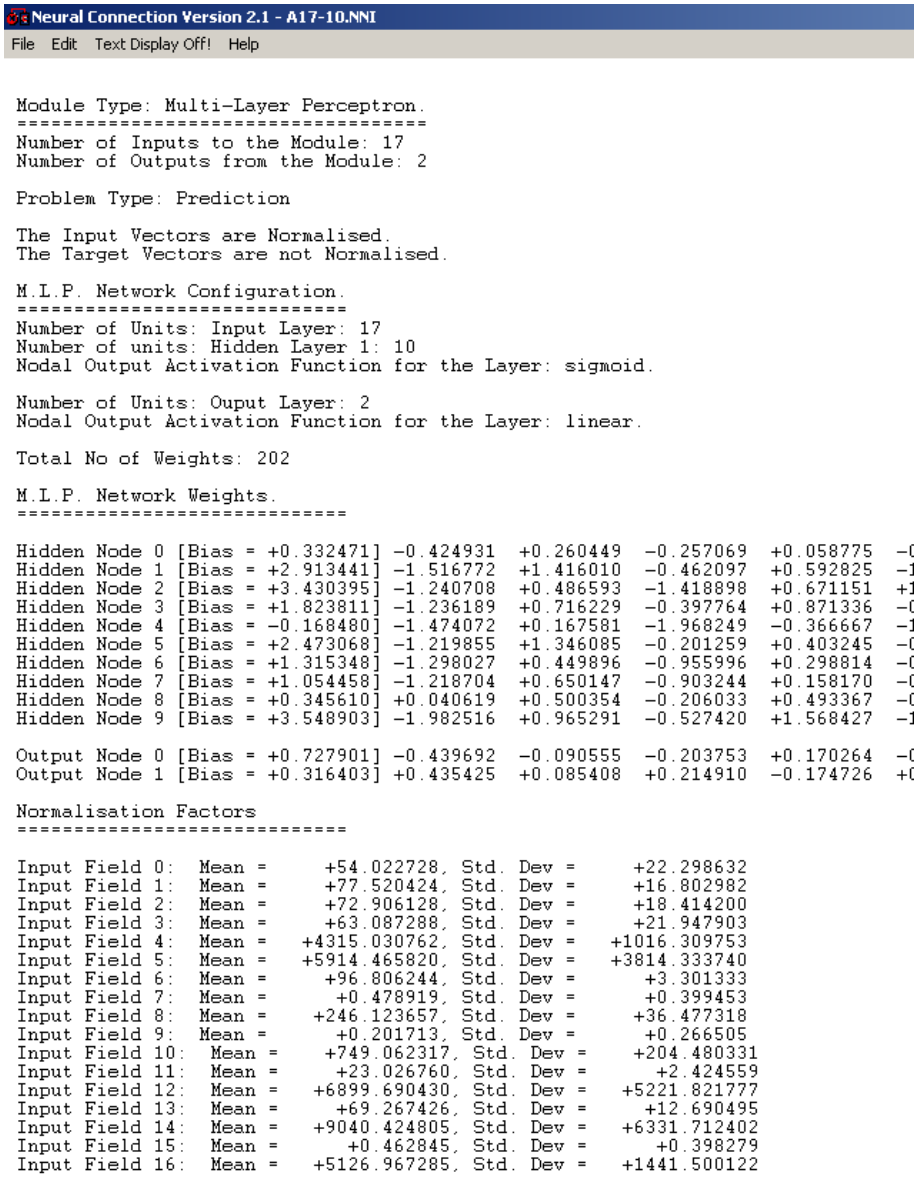
```
clear
clc
%
%This command calls the M-file that holds the 141 Failure observations
%It should be in the same folder as this code to execute properly
%SEE the M-file for a list of variables in the data set

FailsData

n=163;
BOOT=zeros(n,size(fails,2));
f=length(fails);
x=1/f;
failnum=0:1:f;
interval=0:x:1;
intervals=[failnum' interval'];
obs=zeros(f,1);

random=rand(n,1);

for bootnum=1:n
   %index=(find(random(bootnum)>=intervals(:,2)));
   %result(bootnum)=intervals(length(index));
   k=ceil(random(bootnum)*f);
   obs(k,1)=obs(k)+1;
   BOOT(bootnum,:)=fails(k,:);
end
save BootstrapData.out BOOT –ASCII
```

Appendix B. Matlab Code for BAT Equating Table Application

```
clear
clc

%The "data.m" file contains the data to be transformed in the following format
% input=[SSN    EQMPOT     ITMR      TMSD     AIAP      AIAR];

data

%The look-up table is held here in the code, but could also be held in a separate .m file
%This is just a sample of part of one of the look-up tables

lookup=[-1000          -2.1359
         -2.55    -2.1099
         -2.5     -2.0833
         -2.45    -2.056
         -2.4     -2.028
         -2.35    -1.9994];

result=zeros(length(input),1);

for n=1:length(input)
   index=(find(input(n,2)>=lookup(:,1)));
   result(n)=lookup(index(end),2);
   output(n,1)=input(n,1);
   output(n,2)=input(n,2);
   output(n,3)=result(n);
end

save "FILENAME".out output –ASCII
```

Appendix C. Instructions For Preprocessing Network Weights

Access the network weights after training a network in Neural Connections involves several steps. The goal is to reproduce the weights as seen in the figure below. This window can be seen by clicking on the network tool/Status from within a trained network in Neural Connections.



Text File Preprocessing Instructions

1. Open the network architecture from within a .TXT file by clicking File/Open and browsing to the .NNI Neural Connection network file that has been trained and saved.

2. See the Neural Connection User's Guide for help locating the "BEST WEIGHTS." If the standardization option was selected from within the network tool dialog box, them the feature means, standard deviations should also be located at "INPUT NORM."

3. Delete all other data including the headings "BEST WEIGHTS" and "INPUT NORM" leaving only the actual data. Note that the output weights are simply appended to the end of "BEST Weights" data with no method of identifying them. Once the VBA macro is run, the last two rows of the weights are really the output weights.

4. The vectors of means and standard deviations are separated by a third vector of numbers that is placed in the middle. The author does not know what these numbers represent. Each vector begins with the number of features. Locating this number will aid in adding carriage returns to separate the means from the standard deviations. Delete the unknown numbers that are found between the means and standard deviations.

5. Now you should have three "blocks" of data in the text file; the best weights, the means, and the standard deviations. This is a comma delimited file, so add a comma at the very beginning of the first row of the means vector and standard deviations vector.

6. Every new row of the text file should begin with an under score. The only exception is for the very first data point in the text file & the two commas added in step 5.

7. Save the text file.

8. Open an Excel spreadsheet with the code in Appendix D written as a macro. Set up a sheet named "New" by labeling nodes along the rows and the features along the columns. Begin the nodes with node 0 and the features with a column for the bias. Label the last two rows for the output nodes.

9. Ensure that the file path to the text file is correct in the VBA macro

10. Ensure that the cell where you wish to place the first weight is named "matrixstart"

11. Ensure that the numbers used in several places in the VBA code are correct for your specific number of nodes and variables.

12. Note that Neural Connections automatically adds a bias term and an associated node. Also note that the node count begins at zero, therefore the VBA starts at zero also. When inputting the number of nodes in the VBA code, do not include the bias node. Simply use the number of nodes you specified in the network tool dialog box.

Appendix D. VBA Code For Accessing Network Weights From Neural Connections

```
'This file compiles Neural Network “BEST Weights” from a text file that is comma
‘delimited & has underscores at the beginning of each line
‘The text file must be preprocessed manually to delete everything except the “BEST
‘Weights” and the “Input Norm” data, which holds the means and standard deviations
‘used to standardized the data prior to training the network

Sub NetWts()

Dim countrow As Integer,
Dim countcol As Integer
Dim newcell As String
Dim character As String

newcell = ""
countrow = 0
countcol = 0

‘The file path must be correct and end with the name of the txt file containing the weights

Open "C:\THESIS\MLP Recode FlyAero\weightsA34NP-.txt" For Input As 1

Do While Not EOF(1)
   character = Input(1, #1)
   If character <> "_" Then
   If character <> Chr(13) Then
   If character <> Chr(10) Then
   If character <> "," Then
      newcell = newcell & character

      Else
         Worksheets("New").Range("matrixstart").Offset(countrow, countcol) = newcell
         newcell = ""

'the number here is the number of actual variables not including a bias

         If countcol < 34 Then
            countcol = countcol + 1

            Else
               countcol = 0
               countrow = countrow + 1
```

```
'the number here is the number of nodes + the number of outputs

                If countrow = 36 Then

'the number here is always 2 more to move the means and std devs down in the SS

                    countrow = 38

                End If
              'End Else
            End If

'the number here is the same as the number above for countrow = ??

            If ((countrow > 36) And (countcol = 0)) Then
               countcol = 1
            End If
         'End Else

      End If
      End If
      End If
      End If

Loop

Worksheets("New").Range("matrixstart").Offset(countrow, countcol) = newcell

Close #1
End Sub
```

Appendix E. Mean SNR's For All Features Considered

PCSM Included (N = 8)			PCSM Excluded (N = 5)		
Rank	**Feature**	**Mean SNR**	**Rank**	**Feature**	**Mean SNR**
1	Pilot	15.05	**1**	BAT	21.14
2	BAT	14.62	**2**	Pilot	21.07
3	ROTC	12.02	**3**	Nav	17.39
4	Quant	11.47	**4**	Quant	15.91
5	AERO34	10.88	**5**	AIAR	15.49
6	ITMR	10.46	**6**	PS2Z2	15.26
7	BAT_Age	10.21	**7**	ITMP	15.23
8	PS2Z2	10.14	**8**	ROTC	14.78
9	**PCSM**	9.86	**9**	TMSD	14.12
10	AERO2	9.63	**10**	AERO34	14.00
11	H2CX1	9.21	**11**	ITMR	13.95
12	Instruct	8.97	**12**	BAT_Age	13.38
13	TMSD	8.84	**13**	PS2Y2	12.65
14	PS2Y2	7.94	**14**	AIAP	12.29
15	Nav	7.65	**15**	PS2X2	11.83
16	AIAR	7.51	**16**	AERO2	11.81
17	ITMP	7.46	**17**	H2CX1	11.79
18	PS2X2	6.82	**18**	Instruct	10.59
19	Verbal	5.23	**19**	AD	9.83
20	Ed_Level	4.90	**20**	Ed_Level	9.51
21	FLY57	4.81	**21**	FLY57	8.76
22	FLY89	4.69	**22**	MultiInstrm	8.65
23	AIAP	3.72	**23**	Verbal	8.57
24	Other_Stat	3.54	**24**	Other_Stat	8.41
25	FixSngl	3.49	**25**	ANGAFR	8.41
26	AERO1	3.35	**26**	FixSngl	8.16
27	ANGAFR	3.14	**27**	FLY89	8.12
28	AD	2.98	**28**	OTS_Civ	8.07
29	Acad	2.84	**29**	FLY14	5.84
30	FLY14	2.23	**30**	AERO1	5.28
31	OTS_Civ	1.58	**31**	Acad	5.15
32	MultiInstrm	1.19	**32**	GPA	4.10
33	GPA	0.10	**33**	OTS_AD	3.12
34	OTS_AD	-0.75			

Appendix F: DATA_A Factor Loadings Matrices

Unrotated Factor Loadings

Component Matrix[a]

	Component							
	1	2	3	4	5	6	7	8
PILOT	.527	.562	.239	.107	-.012	.103	.069	.134
BAT	.140	.582	-.635	.002	-.009	-.397	.139	.056
FLTHR_CD	.851	-.232	.045	.002	-.247	.010	.030	.022
NAV	.259	.772	.324	.176	.015	.141	.041	.169
QUANT	.070	.753	.400	.196	.005	.095	.016	.151
VERBAL	.150	.605	.443	.035	.039	-.125	.002	-.041
ACAD	.129	.792	.488	.143	.023	-.012	.012	.069
BAT_AGE	.639	-.131	.086	-.367	.457	-.062	-.006	-.072
ED_LEV_C	.749	-.066	.130	-.312	.376	-.057	-.060	-.126
AERO_CD	.824	-.251	.007	.165	-.273	-.015	.031	.045
H2CX1	-.260	-.338	.471	.110	-.035	-.100	.063	-.018
PS2X2	-.235	-.415	.609	.007	-.025	-.178	.095	.034
PS2Y2	-.241	-.380	.573	.033	.027	-.199	.120	.063
PS2Z2	-.206	-.373	.580	.013	-.042	-.201	.081	.041
ITMR	-.024	-.393	.223	.032	.041	.401	-.049	.018
ITMP	.237	.035	-.074	-.054	.175	.204	.103	.097
TMSD	.163	.348	-.462	.032	-.017	.007	.101	.017
AIAP	.061	.133	.170	-.036	-.226	.160	-.633	-.350
AIAR	.134	-.303	-.034	.006	.175	.707	.188	.109
GPA	.059	.105	.097	.218	.155	.007	.367	-.261
ROTC	-.801	.051	-.131	.253	-.327	.077	-.016	.088
OTS_AD	.080	.024	-.009	-.009	.129	.000	.293	-.107
OTS_C	.299	.040	.032	.267	.066	.168	.155	-.601
AD	.365	.260	.217	-.497	.024	-.040	-.301	-.092
ANG	.300	-.328	-.073	-.060	.339	-.145	-.176	.521
AFR	.168	-.141	-.027	-.006	.044	-.211	.361	-.180
OTH_STAT	.115	-.035	.010	.030	-.033	.231	.136	.289
FIXED	.752	-.159	.046	-.213	-.512	-.009	.144	.055
SINGLE	.754	-.161	.050	-.213	-.510	-.008	.145	.054
MULTI	.523	-.251	-.070	.606	.117	-.117	-.187	-.002
INSTRUM	.563	-.279	-.067	.537	.055	-.079	-.154	.029
INSTRUCT	.420	-.236	-.070	.598	.126	-.096	-.163	.022

Extraction Method: Principal Component Analysis.

a. 8 components extracted.

Varimax Rotates Factor Loadings

Rotated Component Matrix[a]

	Component							
	1	2	3	4	5	6	7	8
PILOT	.720	-.215	.263	.190	.120	.081	.063	-.016
BAT	.134	-.772	.003	-.048	-.018	-.482	.142	-.270
FLTHR_CD	.011	-.045	.758	.334	.371	.114	.045	.028
NAV	.899	-.190	.030	.020	.008	.060	.013	-.002
QUANT	.884	-.067	-.083	-.077	-.048	.000	-.015	.017
VERBAL	.724	.047	-.012	.130	-.074	-.206	.067	.095
ACAD	.939	-.015	-.055	.024	-.066	-.114	.029	.063
BAT_AGE	-.024	-.029	.168	.847	.098	.078	.141	-.054
ED_LEV_C	.070	-.052	.262	.848	.166	.051	.126	.045
AERO_CD	.008	-.049	.735	.216	.496	.100	.057	.006
H2CX1	-.051	.637	-.061	-.129	.016	-.009	.044	.008
PS2X2	-.051	.793	-.002	-.034	-.040	-.039	.036	-.064
PS2Y2	-.033	.752	-.048	-.033	-.021	-.048	.061	-.117
PS2Z2	-.026	.746	.017	-.033	-.026	-.070	.024	-.069
ITMR	-.157	.318	.009	.017	.063	.468	-.074	.127
ITMP	.058	-.166	.047	.199	.035	.262	.079	-.099
TMSD	.073	-.584	.046	-.048	.013	-.064	.101	-.084
AIAP	.124	.005	.038	.069	.039	-.083	-.415	.670
AIAR	-.152	-.015	.034	.067	.030	.805	.100	-.002
GPA	.155	.027	-.069	-.026	.082	.027	.506	.068
ROTC	-.112	.061	-.335	-.801	-.208	-.050	-.176	.009
OTS_AD	.018	-.031	.007	.077	-.037	.043	.328	-.055
OTS_C	.069	-.088	.073	.078	.248	.085	.496	.493
AD	.255	-.064	.207	.556	-.234	-.132	-.241	.213
ANG	-.167	.050	.017	.407	.320	.108	-.285	-.503
AFR	-.128	.051	.161	.098	.040	-.126	.423	-.084
OTH_STAT	.079	-.026	.138	-.046	.020	.314	-.035	-.210
FIXED	.012	-.052	.934	.207	.068	.044	.014	-.001
SINGLE	.012	-.048	.934	.209	.069	.046	.016	.000
MULTI	-.018	-.025	.144	.100	.858	.012	.046	.022
INSTRUM	-.037	-.024	.237	.110	.804	.056	.032	.005
INSTRUCT	-.022	-.011	.077	.043	.797	.028	.042	-.004

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 11 iterations.

Appendix G: Data Provided to RANGE J for Correlation Correction

	Unrestricted Population		Restricted Sample	
Variable	Mean	Std Dev	Mean	Std Dev
pcsm	48.64	32.84	63.44	30.87
pilot	70.28	19.86	78.41	16.37
bat	51.34	23.42	54.75	22.12
flthrcd	3.09	3.45	4.52	3.57
nav	66.79	21.06	73.6	18.28
hc2x1	5226.93	1624.12	5081.86	1429.85
ps2z2	6214.25	5211.62	5794.87	3635.89
itmr	748.36	216.84	745.62	205.48
tmsd	237.7	39.04	247.24	35.66
aiap	69.03	13.21	69.28	12.73
pass/fail	Unknown	Unknown	0.94	0.23

Uncorrected Correlations in the Sample											
	pcsm	pilot	bat	flthrcd	nav	hc2x1	ps2z2	itmr	tmsd	aiap	pass
pcsm	1	0.668	0.580	0.690	0.438	-0.367	-0.367	-0.254	0.387	-0.053	0.189
pilot	0.698	1	0.217	0.345	0.828	-0.200	-0.162	-0.145	0.170	0.097	0.180
bat	0.630	0.300	1	-0.035	0.233	-0.441	-0.498	-0.469	0.577	-0.237	0.080
flthrcd	0.677	0.340	0.024	1	0.053	-0.097	-0.057	0.081	0.039	0.048	0.143
nav	0.507	0.856	0.299	0.086	1	-0.174	-0.144	-0.191	0.159	0.117	0.128
hc2x1	-0.374	-0.243	-0.455	-0.120	-0.188	1	0.327	0.184	-0.388	-0.066	-0.083
ps2z2	-0.350	-0.234	-0.492	-0.074	-0.206	0.376	1	0.180	-0.245	-0.021	-0.089
itmr	-0.265	-0.184	-0.453	0.045	-0.213	0.184	0.185	1	-0.140	-0.015	-0.047
tmsd	0.453	-0.253	0.600	0.103	0.223	-0.408	-0.271	-0.149	1	0.032	0.080
aiap	-0.020	0.125	-0.151	0.043	0.118	-0.108	-0.095	-0.063	0.099	1	0.033
pass	0.219	0.201	0.119	0.154	0.169	-0.112	-0.131	-0.068	0.028	0.050	1
Correlations Corrected to Estimate the Population											

Bibliography

AETC. *A Brief Description of the Operational Form of the Basic Attributes Test and Computation of PCSM Scores*. Randolph AFB TX: Air Education and Training Command Studies and Analysis Squadron, 1998.

AETC Instruction 36-2205. *Formal Aircrew Training Administration and Management*. Maxwell Air Force Base AL: Headquarters Air Force Education Training Command, June 2001.

AFI 65-503, Attachment A34-1. *Representative Aircrew Training Costs*. Washington DC: Secretary of the Air Force Financial Management, September 2001.

Air Training Command, Attachment 1. *Pilot Candidate Selection Method (PCSM)*: *Program Guidance Letter*. Randolph AFB TX: Headquarters Air Training Command, April 1985.

Arth, T.O. and others. *Air Force Officer Qualifying Testing (AFOQT): Predictors of undergraduate pilot training and undergraduate navigator training success: Report Number AFHRL-TP-89-52*. Air Force Human Resources Laboratory, Manpower and Personnel Research Division, Brooks Air Force Base TX, 1990. (AD-A221-674).

Bauer, K. W. Class Notes, OPER 685, Applied Multivariate Data Analysis. School of Engineering and Management, Air Force Institute of Technology, Wright-Patterson AFB OH, Spring 2002a.

Bauer, K. W. Class Notes, OPER785, Applied Multivariate Data Analysis II. School of Engineering and Management, Air Force Institute of Technology, Wright-Patterson AFB OH, Fall 2002b.

Bauer, K.W., Alsing, S. G., & Greene, K. A. "Feature Screening using Signal-to-Noise Ratios," *Neurocomputing*, 31: 29-44, 2000.

Belue, L.M. and Bauer K.W. "Determining Input Features for Multilayer Perceptrons," *Neurocomputing*, 7:111-121, 1995.

Birnbaum, Z.W., Paulson, E., & Andrews, F.C. "On the Effect of Selection Performed on Some Coordinates of a Multi-Dimensional Population," Psychometrika, 15(2): 191-204, June 1950.

Bobko, P. "An analysis of Correlations Corrected for Attenuation and Range Restriction," *Journal of Applied Psychology*, 68(4): 584-589, 1983.

Bryant, N.D., "Correcting Correlations for Restrictions in Range Due to Selection on an Unmeasured Variable," *Educational and Psychological Measurement*, 32: 305-310, 1972.

Carretta, T.R. *Time-Sharing Ability as a Predictor of Flight Training Performance: Interim Technical Paper, January 1983-September 1986.* Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB TX, June 1987 (AD-A181838).

Carretta, T.R. *Spatial Ability as a Predictor of Flight Training Performance: Interim Technical Paper, January 1982-September 1986.* Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB TX, July 1987 (AD-A183141).

Carretta, T.R. *Field Dependence-Independence and Its Relationship to Flight Training Performance: Interim Technical Paper, September 1983-December 1986.* Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB TX, December 1987 (AD-A188888).

Carretta, T.R. *Basic Attributes Test (BAT): A Preliminary Comparison Between Reserve Officer Training Corps (ROTC) and Officer Training School (OTS) Pilot Candidates: Technical Paper, July 1986-August 1989.* Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB TX, March 1990 (AD-A224093).

Carretta, T.R. *Cross-Validation of Experimental USAF Pilot Training Performance Models: Interim Technical Paper, July 1986-August 1989.* Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB TX, May 1990 (AD-A222253).

Carretta, T.R. *Short-Term Test-Retest Reliability of an Experimental Version of the Basic Attributes Test Battery: Interim Technical Paper, March 1989-February 1991.* Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Division, Brooks AFB TX, June 1991 (AD-A237484).

Carretta, T.R. *Comparison of Experimental U.S. Air Force And Euro-NATO Pilot Candidate Selection Test Batteries: Interim Technical Paper, 21 May 1991-04 June 1991.* Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Division, Brooks AFB TX, October 1991 (AD-A242358).

Carretta, T. R. "Recent Developments in U. S. Air Force Pilot Candidate Selection and Classification," *Aviation, Space, and Environmental Medicine*, 63, 1992a.

Carretta, T.R. "Understanding the Relations Between Selection Factors and Pilot Training Performance: Does the Criterion Make a Difference?," *The International Journal of Aviation Psychology*, 2(2): 95-105, 1992b.

Carretta, T.R. "Group Differences on U.S. Air Force Pilot Selection Tests," *International Journal of Selection and Assessment*, 5, 1997.

Carretta, T.R. "U.S. Air Force Pilot Selection and Training Methods," *Aviation, Space, and Environmental Medicine*, 71(9), 2000.

Carretta, T.R. & Ree, M.J. *Basic Attributes Test (BAT): Operational Pre-Implementation Analysis and Score Equating: Report Number AL-TP-1993-0015*. Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division, Brooks Air Force Base TX. 1993a.

Carretta, T.R. & Ree, M.J. *Pilot Candidate Selection Method (PCSM): What Makes it Work?: Report Number AL-TP-1992-0063*. Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division, Brooks Air Force Base TX. 1993b.

Carretta, T. R., & Ree, M. J. "Pilot Candidate Selection Method Sources of Validity," *The International Journal of Aviation Psychology*, 4(2): 103-117, 1994.

Carretta, T.R. & Ree, M.J. "Air Force Officer Qualifying Test Validity for Predicting Pilot Training Performance," *Journal of Business Psychology*, 9, 1995a.

Carretta, T.R. & Ree, M.J. "Near Identity of Cognitive Structure in Sex and Ethnic Groups," Personality and Individual Differences, 1995b: 149-155.

Carretta, T.R. & Ree, M.J. "Factor Structure of the Air Force Qualifying Test: Analysis and Comparison," *Military Psychology*, 8: 29-42, 1996.

Carretta, T.R., & Ree, M.J. *Lack of Ability is Not Always the Reason for High Attrition*. Air Force Research Laboratory, Human Effectiveness Directorate, Brooks Air Force Base TX, 1997.

Carretta, T.R., & Ree, M.J. *Pilot Selection Methods: Tech. Rep. No. AFRL-HE-WP-TR-2000-0116*. Human Effectiveness Directorate, Crew System Interface Division, Wright-Patterson AFB OH. 2000a.

Carretta, T.R., & Ree, M.J. "U.S. Air Force Pilot Selection and Training Methods," *Aviation, Space, and Environmental Medicine*, 71(9): 950-956, September 2000b.

Carretta, T.R. & Siem, F.M. *Personality, Attitudes, and Pilot Training Performance: Final Analysis: Technical Paper, September 1983-December 1987.* Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB TX, October 1988 (AD-A199983).

Cohen, J., "The Cost of Dichotomization," *Applied Psychological Measurement*, 7(3): 249-253, Summer 1983.

Cronbach, L.J. "Test Validation," In Thordike, R.L. (Ed), *Educational Measurement*, 2nd edition, pp. 443-507, American Council on Education, Washington DC, 1971.

Damos, D.L. "Pilot Selection Batteries: Shortcomings and Perspectives," *The International Journal of Aviation Psychology*, 6(2): 199-209, 1996.

Dillon, W. R. and Goldstein, M. *Multivariate Analysis, Methods and Applications*. John Wiley & Sons. New York, 1984.

Duan & others. "The Accuracy of Different Methods for Estimating the Standard Error of Correlations Corrected for Range Restriction," *Educational & Psychological Measurement*, 57(2), April 1997.

Earles, J.A., & Ree, M.J. "The Predictive Validity of the ASVAB for Training Grades," *Educational and Psychological Measurement, pp. 721-725, 1992.*

Forsyth, R.A. "*An Empirical Note on a Correlation Coefficient corrected for restriction in* Range," *Educational and Psychological Measurement*, 31: 115-123, 1971.

Greener, J.M. & Osburn, H.G. "An Empirical Study of the Accuracy of Corrections for Restriction in Range Due to Explicit Selection," *Educational and Psychological* Measurement, 3(1): 31-41, 1979.

Greener, J.M. & Osburn, H.G. "Accuracy of Corrections for Restriction in Range Due to Explicit Selection in Heteroscedastic and Nonlinear Distributions," *Educational and Psychological Measurement*, 40(2): 337-346, Summer 1980.

Gross, A.L., & Fleischman, L. "Restriction of Range Corrections When Both Distribution and Selections Assumptions are Violated," *Applied Psychological Measurement*, 7(2): 227-237, Spring 1983.

Hanges, P.J. & Rentsch, J.R. "Determining the Appropriate Correction When the Type of Range Restriction is Unknown: Developing a Sample-Based Procedure," *Educational & Psychological Measurement*, 51(2), Summer 1991.

Held, J.D. "Explanations for Sign Changes in Correcting for Range Restriction." Navy Personnel Research and Development Center, San Diego, CA (1996), http://www.ijoa.org/imta96/paper51.html, 18 October 2002.

Hermelin, E. & Robertson, I.T. "A Critique and Standardization of Meta-Analytic Validity Coefficients in Personnel Selection," *Journal of Occupational and Organizational Psychology*, 74: 253-277, 2001.

Hosmer, D.W. & Leveshow, S. *Applied Logistic Regression*. John Wiley & Sons. New York, 1989.

Johnson, J. T., & Ree, M. J. "RANGEJ: A Pascal Program to Compute the Multivariate Correction for Range Restriction," *Educational and Psychological Measurement*, 54: 693-695, 1994.

Johansson, E.M., Dowla, F.U., & Goodman, D.M. "Back propagation Learning for Multilayer Feed-Forward Neural Networks Using the Conjugate Gradient Method," *International Journal of Neural Systems*, 2(4): 291-301.

Jones, G.E. & Ree, M.J. "Aptitude Test Score Validity: No Moderating Effect Due to Job Ability Requirement Differences," *Educational and Psychological Measurement*, 58(2): 284-294, April 1998.

Kennedy, E. "Estimation of the Squared Cross-Validity Coefficient in the Context of Best Subset Regression," *Applied Psychology Measurement*, 12(3): 231-237, 1988.

Lee, R. & Foley, P.P. "Is the Validity of a Test Constant Throughout the Test Score Range?," *Journal of Applied Psychology*, 71(4): 641-644, 1986.

Lee, R., Miller, K.J., & Graham, W.K. "Corrections for Restriction of Range and Attenuation in Criterion-Related Validation Studies," *Journal of Applied Psychology*, 67(5): 637-639, 1982.

Li, G. & others, "Acceleration of Back Propagations Through Initial Weight Pre-Training with Delta Rule," *IEEE,* Departments of Electrical Engineering and Computer Science, University of British Columbia, Vancouver, B.C., 1993.

Linn, R.L., Harnisch, D.L., & Dunbar, S.B. "Correcting for Range Restriction: An Empirical Investigation of Conditions Resulting in Conservative Corrections," *Journal of Applied Psychology*, 66(6): 655-663, 1981.

Looney, C. G. *Pattern Recognition Using Neural Networks*. Oxford University Press, New York, 1997.

Lord, F.M. & Novick, M.R. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Reading MA, 1968

Mendoza, J.L., Hart, D.E., & Powell, A. "A Bootstrap Confidence Interval Based on a Correlation Corrected for Range Restriction," *Multivariate Behavior Research*, 26(2): 255-269, 1991.

Ness, G. D. *Pilot Candidate Selection Method (PCSM) Evaluation Study: Report Number 96-03S*. Air Education and Training Command Studies and Analysis Squadron, Randolph AFB TX, 1996.

Olea, M.M. & Ree, M.J. "Predicting Pilot and Navigator Criteria: Not Much More Than *g*," *Journal of Applied Psychology*, 79:845-851, 1994.

Perrone, M. P., and Cooper, L. N. *When Networks Disagree: Ensemble Methods for Hybrid Neural Networks*. Brown University, RI, October 27, 1992.

Pugh, D. Chief, Modeling & Simulation, Command Studies Flight, AETC Studies and Analysis Squadron, Randolph AFB TX. Personal Correspondence, January 2003.

Raju, N.S. and others. "A Logistic Regression Model for Personnel Selection," *Applied Psychological Measurement*, 15(2): 139-152, June 1991.

Ree, M.J & Carretta, T.R. *The Correlation of Cognitive and Psychomotor Tests: Interim Technical Paper, January1991-June 1991.* Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Division, Brooks AFB TX, September 1992 (AD-A256413).

Ree, M.J & Carretta, T.R. *Correlation of General Cognitive Ability and Psychomotor Tracking Tests: Final Technical Paper, October 1994-March 1995.* Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Division, Brooks AFB, TX, August 1995 (AD-A297608).

Ree, M.J & Carretta, T.R. *Interchangeability of Verbal and Quantitative Scores for Personnel Selection: An Example.* Air Force Research Laboratory, Human Resources Directorate, Warfighter Training Research Division, Brooks AFB TX, September 1998 (AD-A354026).

Ree, M.J. & Carretta, T.R. *The Role of Measurement Error in Familiar Statistics*. Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, Wright-Patterson AFB OH, Unpublished Report, January 2003.

Ree, M.J., Carretta, T.R., & Earles, J.A. "In Top-Down Decisions, Weighting Variables Does Not Matter: A Consequence of Wilks' Theorem," *Organizational Research Methods*, 1(4): 407-420, October 1998.

Ree, M.J., Carretta, T.R., & Earles, J.A. "In Validation Sometimes Two Sexes Are One Too Many: A Tutorial," *Human Performance*, 12(1): 79-88, 1999a.

Ree, M.J., Carretta, T.R., & Earles, J.A. *Salvaging Construct Equivalence Through Equating.* Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, Wright-Patterson AFB OH, August 1999b (AD-A372382).

Ree, M.J. and others. "Sign Changes When Correcting for Range Restriction: A Note on Pearson's and Lawley's Selection Formulas," *Journal of Applied Psychology*, 79(2): 298-301, 1994.

Rennie, K.M. *Exploratory and Confirmatory Rotation Strategies in Exploratory Factor Analysis*. Presented at the annual meeting of the Southwest Educational Research Association, Austin TX, January 1997, http://ericae.net/ft/tamu/Rota.htm , 18 January 2003.

Sackett, P.R. & Wade, B.E. "On the Feasibility of Criterion-Related Validity: The Effects of Range Restriction Assumptions on Needed Sample Size," *Journal of Applied Psychology*, 68(3): 374-381, 1983.

Sackett, P.R. & Yang, H. "Correction for Range Restriction: An Expanded Typology," *Journal of Applied Psychology*, 85(1): 112-118, 2000.

Schmidt, F.L., Hunter, J.E., & Urry, V.W. "Statistical Power in Criterion-Related Validation Studies," *Journal of Applied Psychology*, 61(4): 473-485, 1976.

Skinner, J. & Ree, M. J. *Air Force Officer Qualifying Test (AFOQT): Item and Factor Analysis of Form 0: Report Number AFHRL-TR-86-68*, Air Force Human Resources Laboratory, Manpower and Personnel Division, Brooks AFB TX, 1987 (AD-A184 975).

Smith, M. *Neural Networks for Statistical Modeling*. International Thomson Computer Press, Boston MA, 1996.

SPSS. *Neural Connection 2.0 User's Guide*. SPSS Inc. and Recognition Systems Inc. Chicago IL, 1997.

Stanley, J.C. Reliability. In Thordike, R.L. (Ed), *Educational Measurement*, 2nd edition, pp. 356-442, American Council on Education, Washington DC, 1971.

StatSoft Incorporated. "Electronic Textbook," http://www.statsoftinc.com/textbook/stdiscan.html , 18 January 2003.

Steppe, J.M. & Bauer, K.W. "Improved Feature Screening in Feedforward Neural Networks," *Neurocomputing*, 13:47-58, 1996.

Waldman, D.A. & Avolio, B.J. "Homogeneity of Test Validity," *Journal of Applied Psychology*, 74(2): 371-374, 1989.

Weeks, J.L. *USAF Pilot Selection*. Proceedings of the First Annual International Air Power Training Conference, London, UK; 2-3 April 1998.

Weeks, J.L., and Zelenski, W.E. *Entry to USAF Undergraduate Flying Training. Tech. Rep. No. AFRL-HE-AZ-TR-1998-0077*. Air Force Research Laboratory, Training Effectiveness Branch, Warfighter Training Research Division. Brooks AFB TX, 1998.

Whitaker, J.S. *Use of Stepwise Methodology in Discriminant Analysis*. Presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 1997, http://ericae.net/ft/tamu/STEPWIS.htm , 18 January 2003.

Williams, H. P., Albert, A. O., & Blower, D. J. *Selection of Officers for U. S. Naval Aviation Training*. NAS Pensacola FL: Naval Aerospace Medical Research Laboratory, November 1999.

Young, I. A. *Development of a Pilot Candidate Selection Model Using Multivariate Techniques*. MS Thesis, AFIT/GOR/ENS/02-18. School of Operational Science, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2002.

Young, I.A, Bauer, K.W., Chambal, S.P., & Pugh, D.M. "Fusing Statistical and Neural Classification for Screening Undergraduate Pilot Training Candidates." *Military Operations Research,* accepted for publication, 23 December 2002.

Zimmerman, D.W. & Williams, R.H. "Restriction of Range and Correlation in Outlier-Prone Distributions," *Applied Psychological Measurement*, 24(3): 267-280, September 2000.

Vita

Captain Ross A. Keener graduated from Timken Senior High School in Canton, Ohio in 1990. He entered undergraduate studies at Malone College in Canton, Ohio where he graduated with a Bachelor of Arts degree in Secondary Education, majoring in mathematics in 1995. He then taught algebra at Louisville High School in Louisville, Ohio until entering the USAF in August 1998. He was commissioned through Officer Training School at Maxwell AFB, AL on 13 November 1998.

His first assignment was at Kirtland AFB, NM working as a program manager of small experimental satellite acquisitions for the then Space and Missile System Center, Test and Evaluation Directorate, DoD Space Test Program (SMC/TELS), which is now USSPACE DET 12. In 2000, he completed a Masters in Education in the School Treasurer program at Ashland University, in Ashland, Ohio, which he began while teaching at Louisville High School. In August 2001, he entered the Graduate School of Engineering and Management, Air Force Institute of Technology. Upon graduation, he will be assigned as an analyst at the Air Force Personnel Center, Randolph AFB, TX.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
25-03-2003	**Master's Thesis**	Mar 2002 – Mar 2003

4. TITLE AND SUBTITLE

USE OF MULTIVARIATE TECHNIQUES TO VALIDATE AND IMPROVE THE CURRENT USAF PILOT CANDIDATE SELECTION MODEL

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Keener, Ross, A., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640, WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GOR/ENS/03-13

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

Air Education and Training Command Studies and Analysis Squadron (AETC/SAS), 151 J Street East, Suite 2, Randolph AFB, TX 78150
e-mail: david.pugh@randolph.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Pilot Candidate Selection Method (PCSM) seeks to ensure the highest possible probability of success at UPT. PCSM applies regression weights to a candidate's Air Force Officer Qualification Test (AFOQT) Pilot composite score, self-reported flying hours, and five Basic Attributes Test (BAT) score composites. PCSM scores range between 0 and 99 and is interpreted as a candidate's probability of passing UPT.

The goal of this study is to apply multivariate data analysis techniques to validate PCSM and determine appropriate changes to the model's weights. Performance of the updated weights is compared to the current PCSM model via Receiver Operating Curves (ROC). In addition, two independent models are developed using multi-layer perceptron neural networks and discriminant analysis. Both linear and logistic regression is used to investigate possible updates to PCSM's current linear regression weights. An independent test set is used to estimate the generalized performance of the regressions and independent models. Validation of the current PCSM model demonstrated in the first phase of this research is enhanced by the fact that PCSM outperforms all other models developed in the research.

15. SUBJECT TERMS

Pilot Candidate Selection Method (PCSM), Undergraduate Pilot Training (UPT), Multivariate Analysis, Validation, Factor Analysis, Neural Networks, Discriminant Analysis, Linear Regression, Logistic Regression, Receiver Operating Curve (ROC)

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kenneth W. Bauer, Ph.D. (ENS)
U	U	U	UU	234	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-6565, ext 4328; e-mail: Kenneth.Bauer@afit.edu