PREDICTIVE RELATIONS BETWEEN COGNITIVE ABILITIES AND PILOT PERFORMANCE: A STRUCTURAL EQUATION MODELING APPROACH

by

Khalid Saif ALMamari

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Educational Studies

West Lafayette, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Anne Traynor, Chair
Department of Educational Studies

Dr. James Greenan
Department of Curriculum and Instruction

Dr. Yukiko Maeda
Department of Educational Studies

Dr. Richard Olenchak
Department of Educational Studies

Approved by:
Dr. Janet Alsup

This dissertation is dedicated to the souls of my father and mother, may God have mercy on them

ACKNOWLEDGMENTS

First and foremost, I am so grateful to the Almighty God for helping and guiding me to complete this research project. “And if you should count the favors of Allah, you could not enumerate them. Indeed, Allah is Forgiving and Merciful” (Quran, surah An Nahl, 16:18). At the end of my study, I would like to offer my appreciation and thanks to the Royal Air Force of Oman for supporting and facilitating my study, and to the Ministry of Higher Education in Oman for sponsoring and funding my program of study. I would also like to express my deepest gratitude to my academic supervisor, Dr. Anne Traynor, for her patience, support, encouragement, and excellent guidance and instruction. I truly would not have gone this far without her constructive feedback, thoughtful comments, insights, and knowledge. I will always be indebted to have had the honor of working with her. I also extend my appreciation to the members of my thesis committee for their input, feedback, and contributions to my work: Dr. Yukiko Maeda, Dr. James Greenan, and Dr. Richard Olenchak. Dr. Maeda’s useful comments and suggestions provided insight that were particularly valuable and productive.

I’m sincerely thankful to all members of my family, brothers, and sisters for their support. I cannot adequately express how much I appreciate their help, encouragement, and prayers. For my boys, Jihad, Albaraa, Yahya, Azzam, Mohamed, and Saif, thank you for your patience and I hope you will realize the importance of what I was doing. Last but certainly not least, my deepest heartfelt appreciation and gratitude, which cannot be fully explained in words, is to my wife. Her unstinting patience, support, and encouragement throughout this process were beyond compare. Without her sacrifices, the completion of this project would not have been possible. I don't think I can ever repay you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFOQT	The Air Force Officer Qualifying Test
AI	Aviation Information subtest
AR	Arithmetic Reasoning subtest
BC	Block Counting subtest
CFA	Confirmatory Factor Analysis
DI	Data Interpretation subtest
EFA	Exploratory Factor Analysis
EM	Electrical Maze subtest
g	Intelligence/general mental ability/general cognitive ability/general factor
GS	General Science subtest
HF	Hidden Figures subtest
IC	Instrument Comprehension subtest
I/O	Industrial/Organizational
MC	Mechanical Comprehension subtest
MK	Math Knowledge subtest
RC	Reading Comprehension subtest
RB	Rotated Blocks subtest
SEM	Structural Equation Modeling
SR	Scale Reading subtest
TR	Table Reading subtest
VA	Verbal Analogies subtest
WK	Word Knowledge subtest

ABSTRACT

A large body of literature suggests that cognitive abilities are important determinants for training and job performance, including flight performance. The associations between measures of ability tests and job performance have been the focus of many empirical studies, resulting in an overall conclusion that general mental ability, *g*, is the main source of prediction, while other narrower abilities have limited power for predicting job performance. Despite the attention given to cognitive ability-flight performance relationships, their associations have not been fully understood at the broad construct level, and most extant literature focused on the relations at the observed scores level. Thus, the present dissertation study was designed to contribute to the progression of this understanding by examining the relations between cognitive abilities and flight training performance, using data from four U.S. Air Force (USAF) pilot samples. For comparison, one navigator and one air battle manager sample were also analyzed. The data were obtained from correlation matrices of prior investigations and analyzed via structural equation modeling (SEM) procedures.

Four studies are reported in the thesis: (1) preliminary study, (2) primary validation study, (3) cross-validation study, and (4) cross-occupation validation study. The preliminary study assessed the test battery used in the subsequent predictive studies. The primary validation study introduced a bifactor predictive SEM model for testing the influence of cognitive abilities in predicting pilot performance. The cross-validation study assessed the consistency of the predictive model suggested in the primary validation study, using three additional pilots' samples. The cross-occupation validation study compared the predictive model using data from three aviation-related occupations (flying, navigation, air battle management). Ability factors were extracted from scores of pilot applicants on the Air Force Officer Qualifying Test (AFOQT), the USAF officers' primary

selection test battery, whereas the flight performance scores were obtained from pilot records during the flight training program.

In addition to the *g* factor, *verbal ability, quantitative ability, spatial ability, perceptual speed ability, and aviation-related acquired knowledge* are the six latent cognitive ability factors investigated in the reported studies. Pilot performance measures were modeled either as observed or latent variables covering ratings of academic and hands-on flying performance in different phases of the training program. The studies of this thesis established that (1) general ability contributes substantially to the prediction models; however, it is not the only important predictor, (2) aviation-related acquired knowledge is the most robust predictor of pilot performance among the abilities examined, with a role even exceeding that of *g*, (3) perceptual speed predicted pilot performance uniquely in several occasions, while verbal, spatial, and quantitative abilities demonstrated trivial incremental validity for hands-on pilot performance beyond that provided by the *g* measure, and (4) the relative importance of cognitive abilities tends to vary across aviation occupations.

CHAPTER 1: INTRODUCTION

In recent years, there has been a growing interest in the predictive relations between cognitive abilities and job performance (e.g., Hunter, 1986; Hunter & Hunter, 1984; Lang et al., 2010; Murphy, 1989; Ones et al., 2012; Ree, Carretta, & Steindl, 2001; Schmidt & Hunter, 1998). Industrial and organizational (I/O) psychologists have attempted to validate cognitive abilities' relation to training and job performance, and found that their correlations typically fall somewhere above .50 (Ones et al., 2012; Schmidt, & Hunter, 2004), and that the general factor of ability accounts for a proportion of total variance in a test battery ranging from 30% to 65% (James & Carretta, 2002; Jensen, 1980, pp. 216). Aptitude ability test batteries are useful sources of information about the applicants' cognitive ability, and similarly, performance and achievement during training programs are useful sources of information about the skills and proficiency achieved by the applicants throughout training. Hence, associating scores of these two broad dimensions represents an effective strategy to understand the role of cognitive abilities in subsequent training and job performance. The predictive relations obtained through such a strategy (i.e., *predictive validity*) is one of two criterion-related validity approaches that also include *concurrent validity* (Cronbach & Meehl, 1955; Shepard, 1993).

Ability test batteries are designed to serve several goals, among which are their effectiveness in predicting desirable outcomes. For example, selection test batteries such as the AFOQT (the U.S Air Force Officer Qualification Test; Carretta & Ree, 1996) and the ASTB (the U.S Navy Aviation Selection Test Battery; Williams et al., 2000) are administered to qualify applicants for pilot job careers and to ensure that they meet the minimum requirements for ability competencies required by flying occupations (Carretta, 2000). Moreover, they serve as a convenient tool to forecast the success of selectees in future training and job performance.

Similarly, scholastic test batteries such as the ACT, SAT, or GRE are applied to aid colleges and universities for assessing the academic proficiency of candidates (Kuncel & Hezlett, 2010; Kuncel et al., 2001; Shepard, 1993; Schneider & Dorans, 1999) and providing means for predicting their achievement and program completion in an objective manner (e.g., Coyle & Pillow, 2008). Gathering pieces of evidence concerning the criterion-related validity of test batteries is advantageous for psychometric investigations and increases the confidence in their practicality. Highlighting the prediction role of test batteries can also be useful in efforts to revise their content and improve their functionality. Accordingly, such investigations help to build bridges between the theoretical and practical aspects of cognitive testing, and feed both with important information relevant for their growth and advancement.

In this thesis, I present four investigations targeting the predictive relations between cognitive abilities and pilot performance during a training program (i.e., *ab-initio* pilots). Correlations among indicators of cognitive abilities and flight performance were sourced from prior investigations of the AFOQT, the main selection battery of the U.S Air Force (USAF) officers. The flying job was chosen for the present investigation because aviation-related jobs, especially flying, are considered a complex class of jobs that require cognitive potential and a high level of abilities (Duke & Ree, 1996; Ree et al., 1995). Therefore, the role of cognitive abilities can be more demonstrable than in less cognitively demanding work environments. The AFOQT was the test battery chosen to source the cognitive data because it is one of the most studied test batteries in the aviation literature, and has a long history of use in selection, supported by well-established psychometric properties (Carretta & Ree, 1996; Drasgow et al., 2010).

One notable observation is that previous validation studies have focused primarily on the predictive relations between ability tests and flight performance. The objective in these studies was mostly collecting evidence for criterion-related validity of individual tests or specific test batteries used in pilots' selection (e.g., Barron & Rose, 2017; Carretta & Ree, 1995; Kock & Schlechter, 2009; Wang et al., 2018; Zierke, 2014). The results of such investigations may be more relevant to particular organizations, and by necessity, have limitations restraining their generalizability to the general population of flight training programs. While other organizations benefit from the findings of such local studies, the benefit may not be maximized. Empirical studies focusing on specific organizations are still necessary and aid the scientific fields with indispensable investigations; however, conclusions of external validity should be made cautiously (i.e., Campbell & Stanley, 1966). To overcome the shortcomings in primary studies, and to provide stronger evidence for the relations between variables (i.e., cognitive abilities and job performance), systematic reviews and meta-analyses are frequently proposed as useful techniques (Rosenthal & DiMatteo, 2000).

In aviation, several meta-analyses have assessed the relations between cognitive abilities and flight performance, some of which have been comprehensive in their scope, while others have been narrower. A number of cognitive abilities were highlighted in these studies as critical for pilot performance, such as mechanical ability, perceptual speed ability, spatial ability, reaction time, and general ability (Hunter & Burke, 1994; Martinussen, 1996). Some composite scores, a common product of test batteries, have also been determined to be useful predictors for pilot performance, such as those loaded with broad abilities of acquired knowledge, general ability, and perceptual processing (ALMamari & Traynor, 2019). At a more applied level, individual tests of some specific abilities were found to possess promising predictive utility for pilot performance such as Instrument Comprehension, Mechanical Principles, and Aviation Information (Martinussen & Torjussen, 1998). For the AFOQT, the main instrument utilized in this thesis, a recent meta-analysis has uncovered the main source of its predictive validity (ALMamari & Traynor, 2020). Results showed that the Pilot composite score, followed by scores from the tests of Instrument Comprehension, Scale Reading, Aviation Information, Table Reading, and Data Interpretation, had the highest mean correlation with the overall criterion of flight performance (mean r = .14–.17).

1.1 Statement of the Problem

In the context of the long-standing research effort to understand the role of cognitive abilities in workplaces (e.g., Hunter, 1986; Hunter & Hunter, 1984; Schmidt & Hunter, 1998), and in an attempt to pursue different research methodologies for the study of the ability-performance relationship, I proposed the current thesis motivated by a number of research enquires. I departed in this work from the traditional research design that typically relates observed scores of ability tests with observed scores of job performance via correlation or regression analyses. The main interest here is not to evaluate specific observed scores of individual tests or a test battery; rather, the goal is to test broad factors of cognitive abilities (e.g., verbal, quantitative, and spatial). Hence, I rely on structural equation modeling (SEM) procedures to achieve the goal of assessing the underlying predictive relations between cognitive abilities and job performance. SEM is a general statistical methodology that subsumes and extends correlation, regression, factor analysis, and path analysis (Schumacker & Lomax, 2010). It allows researchers to detect and test models that account for complex multivariate relationships among observed and latent variables (Marcoulides & Yuan, 2017). It also provides useful procedures for establishing connections between variables and analyzing structural relationships, with an explicit estimate of measurement errors (Ullman & Bentler, 2012). SEM allows multiple, interrelated measures to be associated and their properties estimated simultaneously in a single analysis (Kline, 2015). Since its introduction, SEM has influenced theory construction in many fields as it allows for optimal integration between measurement and substantive theory (Guo et al., 2009). Thus, using SEM models in this thesis can be beneficial and provide different views of ability-performance relationships at the construct level.

To put the intended investigations in a proper context of intelligence and I/O psychology research, the ongoing debate of general versus specific abilities was of primary interest. Intelligence theory has debated for a long time the best organization of human cognitive structure (Carroll, 1997). This debate arises from undebatable observations pertaining to the positive correlations (i.e., manifold) commonly noted among ability tests. On the basis of this psychometric phenomenon, Spearman (1904) started this more-than-century-long argument proposing his two-factor theory of intelligence. He hypothesized that the substantial common variance in cognitive performance is due to one single general ability factor, while specific ability factors have negligible explanatory roles beyond that attributed to general ability. After the two-factor hypothesis of cognitive structure (Spearman, 1904), other models have been posited emphasizing different numbers/levels of broad cognitive factors such Cattell's theory of fluid and crystallized intelligence (1963), Thurstone's seven primary mental abilities (1938), Carroll's three stratum theory (1997), and Vernon's four stratum model of intelligence (1964). Despite the shift of focus in intelligence theory and the numerous models proposed, Spearman's model is still alive and is seen in the daily practices of cognitive testing where the final product is often one overall composite score (e.g., IQ).

The version of this debate that took place in I/O psychology research revolves around whether general mental ability, *g*, or specific abilities are the main contributor to job performance prediction. A strong line of research, supported by extensive empirical studies and meta-analysis reviews, reached a conclusion that *g* is the best stand-alone predictor of job performance, whereas specific abilities have trivial prediction role beyond that provided by the score of *g* (e.g., Hunter, 1986; Ree et al., 1994; Schmidt & Hunter, 2004). On the contrary, an increasingly popular emerging line of research argues that specific abilities can be important in some workplaces but might have been overlooked due to methodological limitations in previous studies (Kell & Lang, 2017; Schneider & Newman, 2015). Examples of such limitations include the theoretical background on which ability-performance relations are modeled (Lang et al., 2010), analytical procedures used in studies (Lang & Kell, 2019), predictor-criterion alignment or the ability–performance compatibility principle (Schneider & Newman, 2015; Wee, 2018), and the specific abilities and occupation investigated (Nye et al., 2020). Some of these limitations, as well as the relative importance of general versus specific abilities, are topics addressed in this thesis.

A review of aviation psychology literature showed that the AFOQT studies have contributed towards establishing specific awareness of ability-performance relations. Thus, obtaining cognitive data for pilot students from such an instrument can be a sensible choice. One unique feature of the AFOQT is that it has multiple tests measuring several domains of cognitive structure (Carretta & Ree, 1996). This is an important characteristic as it ensures the existence of a sufficient number of abilities to investigate and manipulate. Ten to 16 subtests were incorporated in the AFOQT across the most recent six forms (Form O, P, Q, S, and T), tapping five main ability factors: *verbal ability, quantitative ability, spatial ability, perceptual speed ability, and aviation-related acquired knowledge and aptitude* (Carretta & Ree, 1996; Drasgow et al., 2010; Carretta et al., 2016). The AFOQT researchers made a considerable effort to validate it for different flight performance criteria, including pass/fail training, academic achievement, daily flight ratings, check flight ratings, and the exceedance of average flying hours of training. Despite this effort, most predictive validity studies of the instrument have used analytical procedures focused on observed ability-performance predictive relations (e.g., correlation, regression). Such methods may not be sufficient to give inference about the more abstract level of broad ability factors. Thus, in this thesis, several AFOQT datasets were reassessed, with a primary focus on first- and higher-order factors of cognitive abilities (that is, second and third stratum), rather than on the individual tests in the battery. Addressing research questions at the broad constructs of abilities is useful for theory building and for explaining the underlying relations (e.g., Bacharach, 1989).

At the subtest level, the meta-analysis of ALMamari and Traynor (2020) might have satisfied the need for assessment of the likely relations between individual ability tests and measures of flying performance. Although the study made an effort to interpret the results at a broader level, specifically on the basis of the five ability domains previously suggested for the AFOQT's structure, the findings remain limited by the observed bivariate correlation data accumulated in the study, which prohibited drawing a firm conclusion about the second and third stratum of abilities. In order to assess the broad ability constructs, there is a need for studies utilizing statistical procedures that are capable of disclosing the underlying latent factors and linking them with scores indicative of job performance. The SEM-based assessment proposed here may fulfill this need, and provides an approach for assessing predictive relations that have not yet been examined. As noted by Evermann and Tate (2009), "latent variables represent theoretical constructs, and hypothesized regression relationships between them represent hypothesized causal propositions between constructs." (p. 2).

1.2 Purpose of the Study

Ackerman and Beier (2012) reasoned that the dearth of new research on intelligence in I/O psychology is a result of failing to distinguish between *g* and intelligence, as well as the expansive results of validity generalization. They argued that broadening the focus of predictors to include intellectual investment constructs, such as broad and specific job knowledge, can enhance the predictive validity and advance the understanding of individual differences in job performance. The current thesis responds to such calls by analyzing data on a variety of specific abilities and job performance measures in training settings. A central goal in this work is to relate cognitive abilities, as indexed by latent factors, with performance measures, as indexed by either observed or latent variables. Via SEM procedures, the cognitive ability-flight training performance relations were assessed considering multiple perspectives and different research questions. The key research question for the present thesis reads: *To what extent are cognitive abilities related to flight training performance criteria?*

Substantive knowledge about the empirical relations between constructs measured prior to training and later flight performance may contribute to the advancement of pilot selection models. Several important applications of the SEM framework were covered by the studies of this thesis (e.g., bifactor approach), most of which are rarely used in the field of aviation psychology. Hence, a further goal of this thesis was to present a practical demonstration of how SEM models can effectively be used in building connections between predictor and criterion variables.

1.3 Overview of the Studies and Suggested Research Questions

The main goal of this thesis is to assess the cognitive abilities-flight performance relationships through a structural equation modeling approach. Four studies were investigated using different samples from USAF pilot trainee populations. Because the primary interest was to assess as many cognitive abilities and flight performances as could be extracted, a data-driven approach was pursued to make maximal use of the available information. Following is a brief description of the four investigations, and more details will be presented in the succeeding sections.

1.3.1 Preliminary study

The first study in this project was planned to provide evidence for the internal construct validity of the AFOQT using meta-analytic and EFA procedures. As described by Cronbach and Meehl (1955), "construct validation is involved whenever a test is to be interpreted as a measure of some attribute" (p. 282). Internal construct validity expresses how accurately the constructs of a scale are differentiated from one another and to what extent they explain the variance found in the sample. Because EFA is an efficient means for establishing a test's construct validity (Thompson & Daniel, 1996), the AFOQT data in this preliminary study was analyzed using that method. Further, instead of relying on a single data set, this study aggregated and meta-analyzed the intercorrelations among 16 AFOQT subtests. Cheung and Chan's (2005) meta-analytic SEM (MASEM) approach was chosen as a methodology for this investigation. However, only the first stage of the two-stage method was applied in this study. The resulting pooled correlation matrix of subtests was then analyzed by means of exploratory factor analysis (EFA). I found this study an essential introductory examination for the succeeding studies, as it provides support for the modeling choices of cognitive abilities and justifies the selection of certain subtests as indicators of ability factors. What makes this investigation more interesting is the scarcity of EFA studies of AFOQT data. To my knowledge, there exists only one EFA study for AFOQT that was reported more than three decades ago (i.e., Skinner & Ree, 1987). Thus, this study would be a useful addition to AFOQT factor analytic literature. Additionally, this study provided an opportunity to assess the extent to which the five-factor model frequently proposed for AFOQT is a viable solution for the data and whether it is extractable. Although the current version of AFOQT includes only 10 of the 16 subtests, the results remain useful since the currently-used subtests also appeared in former versions. The research questions proposed for this study include the following:

(1) How heterogeneous is the pooled correlation matrix?

(2) When imposing one- to six-factor EFA solutions, what is the most plausible model of the AFOQT data?

(3) What is the content of the subtests of each solution?

(4) Is the five-factor model a superior model for the AFOQT data?

1.3.2 Primary Validation Study

This study attempted to assess the relations of five ability latent factors (verbal, quantitative, spatial, perceptual speed, and aviation-related acquired knowledge), after its plausibility had been verified by EFA, with three pilot performance criteria. The debate on the relative importance of the general factor of ability, *g*, and specific ability factors was a major research question in this study and those that follow. The bifactor model (Reise, 2012) was the main procedure used for modeling the latent factors underlying ability and performance measures. Although bifactor models have seen a growing interest (Rodriguez et al., 2016) and have successfully been utilized in predictive studies (e.g., Gustafsson & Balke, 1993), their heavy use has been in factor analytic studies as a measurement model for test batteries or other instruments. Their usage as part of SEM models, particularly as concerns performance modeling, is not as frequent as might be expected. A unique feature of this model is that it allows for disentangling the *g* effect from specific ability effects. It partitions domain-general from domain-specific variance and allows the unique effects of specific abilities to emerge and manifest (e.g., Zhang et al., 2020). I proposed the following research questions for this study:

(1) What are the most predictive cognitive abilities for pilot performance?

(2) Do specific abilities have incremental validity over that provided by the *g* score?

(3) Does the predictive validity of ability factors vary by performance criterion?

1.3.3 Cross-validation Study

The purpose of this study was to cross-validate, or check the consistency of, the findings resulting from the primary validation study across different samples and pilot performance measures. The re-examination of the results in separate validation samples can be advantageous and provide further evidence for the predictive validity of cognitive abilities for pilot performance. Bokhari and Hubert (2018) demonstrated that applying cross-validation when building predictive models contributes to increasing the reliability and replicability of psychological research. Replicating the findings found in the primary study with independent, multiple samples strengthens the overall conclusion determined by this thesis about the cognitive ability-pilot performance relationships. Therefore, three samples representing USAF pilot trainees were obtained from the correlation matrices of prior AFOQT studies to cross-validate the bifactor predictive model established in the primary study. The analyses of all three datasets were identical to those reported in the primary study.

In two of the three samples, the same five ability factors (verbal, quantitative, spatial, perceptual speed, and aviation-related acquired knowledge), in addition to the *g* factor, were assessed for their effects on three flight performance measures. The available data in the third sample allowed an assessment of four abilities (*g*, verbal, quantitative, and aviation-related acquired knowledge) that might predict three latent performance factors. This cross-validation study will investigate similar research questions to those posited in the primary study. It will particularly focus on whether the findings revealed in the primary study are replicable across samples and pilot performance criteria.

1.3.4 Cross-occupation Validation Study

In order to compare the role of cognitive abilities in flying performance with their role in other aviation-related jobs, I designed this study to understand how influential cognitive abilities are in different aviation occupations, specifically: flying, navigation, and air battle management. Another goal in this study is to assess the interplay of general and specific cognitive abilities in predicting performance in the three aviation jobs. Five latent ability factors (verbal, quantitative, spatial ability, perceptual speed, and aviation-related acquired knowledge), as well as the *g* factor, were associated with latent factors of performance for pilot and navigator samples or observed performance scores for an air battle manager sample. Three separate correlation matrices representing pilots, navigators, and air battle managers were reproduced from prior AFOQT research for the purpose of this comparison. This study also utilized bifactor models to obtain a better separation of the effects of general and specific abilities (e.g., Reise, 2012). Below are the posited research questions in this study:

(1) Which cognitive abilities best predict performance in each of the three occupations (flying, navigation, air battle management)?

(2) Is there any incremental validity of specific factors of abilities above that obtained from the *g* factor?

(3) How do the predictive relations of flying jobs differ from those of navigation and air battle management jobs?

CHAPTER 2: LITERATURE REVIEW

After a brief introduction to the significance of studying pilot performance, this chapter will include a review of the literature in five main areas including (1) determinant of pilot performance, (2) the AFOQT, (3) the relative importance of cognitive abilities, (3) modeling the structure of cognitive abilities, and (4) structural equation modeling (SEM).

2.1 Significance of Pilot Performance Study

Flying is a complex class of job (Duke & Ree, 1996) that requires multiple aptitudes and abilities for its applicants. The training program for flying is an expensive one and, in fact, one of the most expensive training programs (e.g., Callender, 2018; Hampton et al., 2017; Goeters & Maschke, 2004). The initial training costs of a military pilot are estimated to be nearly $800,000 (Stokes & Kite, 1994). Due to the high risk of flying jobs and the potential loss of resources (both equipment and human lives) (Bates et al., 1997), research in pilot selection and performance are increasingly important. Proficiency in flying skills provides more assurance of human safety, and thus, it is crucial to select candidates with the "right stuff" for this job (Katz, 2006) who are capable of being successful pilot candidates and have high potential to perform well in future flying assignments.

Although the advancement in aviation technology has helped to reduce the rate of machine failures, the number of aviation mishaps attributed to human errors is continually increasing (Driskell & Olmstead, 1989). For example, Yacavone (1993) indicated that 58% of 308 total mishaps of Class A in the U.S. Navy between 1986 and 1990 resulted from aircrew error. Similarly, Shappell and Weigmann's (1996) study that assessed the U.S. Naval aircraft mishaps from 1977 to 1992 showed that the ratio mishaps due to human error against those due to other factors (e.g., mechanical, environmental) was estimated to be from 1:1 to 9:1 for single-piloted and 12:1 for dual-piloted aircraft, which meant that roughly 150 mishaps every year were due to human error. According to some aviation investigations, human error was found to be responsible for between 60–80% of aviation accidents (Yazgan et al., 2017). Consequently, any improvement in the quality of pilot candidates achieved through enhancing the effectiveness of selection procedures, including test batteries, is typically associated with significant cost savings and reduction in injuries/loss of life. One possible way to accomplish this goal is to ensure that the cognitive ability factors underlying a selection test battery are important to the flying job, and can be predictive for future training performance.

2.2 Determinant of Pilot Performance

2.2.1 Knowledge, Skills, Abilities, And Other Characteristics (KSAOs) of Flying Job

To be a pilot, you need to have a broad array of aptitudes and skills. Training programs are generally designed to invest the already-available cognitive abilities for learning and flying practice. Several job analysis surveys, narrative reviews, and meta-analyses highlighted specific skills and abilities that are deemed essential for aircraft piloting. The variables emphasized in studies tend to vary, although they share many commonalities. Job analysis studies can be a useful source for understanding the knowledge, skills, abilities, and other personal characteristics (KSAOs) applicable for flying. Ideally, the outcome of this analysis is translated to practice informed by cumulative knowledge (Carretta & Ree, 2003). However, Damos (2011) noted that the application of KSAOs for the pilot job might be hindered due to the absence of a unified taxonomy that provides a clear identification and definition of the most promising attributes. Despite this limitation, she was able in her review to determine the importance of perceptual speed and spatial orientation for pilots and, to a smaller and less certain extent, the abilities of numerical/quantitative abilities, multi-tasking attributes, multi-limb coordination, and selective attention. A firm conclusion about mechanical aptitude and situational awareness, two highly-regarded abilities in aviation, was not possible due to differences in their definitions across studies.

Of many approaches proposed for job analyses, the *job analysis survey* is probably the most widely-used approach for assessing jobs' requirements. The simple idea behind this method is that each job is best described and judged by those who have professional involvement in that job (subject matter experts), and thus, they are the best analysts and examiners of the attributes and skills required by that job. The Job Analysis Survey (F-JAS) of Fleishman (1992) is a good illustrative example due to its wide applications. The F-JAS contains 72 skill and knowledge scales distributed across four broad areas that are assessed through the survey. The F-JAS measures the following abilities, with numbers of scales denoted in brackets: Aptitudes (i.e., cognitive abilities [21], psychomotor abilities [10], physical abilities [9], and sensory abilities [12]), knowledge and skills [11], and interactive/social components [9]. Each measure is rated with a 7-point scale evaluating the importance of the construct for successful job performance, according to the respondents' judgment.

In a study on ability requirements for fast jet pilot training in the North Atlantic Treaty Organization (NATO) countries, Carretta et al., (1996) designed a modified version of the Fleishman job analysis survey containing 12 critical tasks specific to the job of a fighter pilot, and distributed it to a pilots sample representing several NATO countries. The abilities that received the highest rating were situational awareness, memory, motivation, and reasoning, while the abilities of reading comprehension, writing, and leadership received the least rating of importance.

Goeters et al. (2004) administered the Fleishman's F-JAS on 141 pilots serving a major European airline (Lufthansa) to establish a general profile of job demands for airline pilots.

Concerning cognitive abilities, time-sharing and spatial orientation were judged the most relevant constructs for successful flying performance. Twelve abilities were rated second in relative importance (i.e., selective attention, perceptual speed, number facility, memorization, and visualization), and six were rated third (oral comprehension, inductive reasoning, mathematical reasoning, and category flexibility). Among psychomotor abilities, constructs of rate control received the highest rating, followed by control precision, response orientation, multi-limb co-ordination, and reaction time. Other attributes that were rated "very relevant" included map reading from the knowledge & skills domain and stress resistance, cooperation, communication, and decision making from the interactive/social skills domain. A U.S. Army aviators job analysis (Kubisiak & Katz, 2006) showed that situational awareness, operation and maneuvering of helicopter, psychomotor ability, information processing, and decision making are the highest-rated attributes by pilots and that tasks associated with emergency procedures and safety had the highest ratings type of tasks.

Job analysis surveys are helpful to show the most vital skills and abilities required by individual occupations. The judgement of experienced and knowledgeable workers in each field is a useful source of information regarding the relative importance of specific skills and abilities. The application of such surveys in aviation revealed strong connections between flying skills and several cognitive and psychomotor ability constructs. In particular, the abilities of information processing, perceptual speed, spatial orientation, time-sharing, and decision making are often rated highly by subject matter experts. Memory constructs have also received high ratings, which may raise a question regarding the lack of studies devoted to this construct in aviation, given this evidence from job surveys that support its role in flying.

2.2.2 Narrative Review of Pilot Performance

A number of articles have reviewed the determinants of pilot performance in attempts to highlight the most important predictors for successful performance. Such revisions informed the qualitative understanding of abilities and attributes considered crucial for pilots. For example, researchers of the recent U.S Army project to introduce a new selection battery for aviators (Paullin et al., 2006) emphasized a dominant role for general intelligence in the prediction of aviator performance, and a possible role for six particular constructs, including psychomotor skills, selective and divided attention, working memory, aviation interest/knowledge, flying experience, and personality, with potential incremental validity beyond that from a measure of *g*. In their review of pilot performance variables, Bates et al. (1997) listed five broad constructs as variables worth assessing in training performance, including general cognitive ability, psychomotor coordination, biographical information, information processing, and personality. Olson et al. (2009) highlighted the role of cognitive ability tests, psychomotor tests, and personality tests as value-added variables for the prediction of pilot performance. Carretta and Ree (2003) stressed that general cognitive ability is the mainstay of military testing and that measures of pilot job knowledge and psychomotor ability have incremental validity above measures of *g*. They suggested that any selection systems should incorporate at least three common measurement components: cognitive ability, conscientiousness, and job knowledge. Street et al. (1992) discussed a set of five outstanding quality and skill domains that were identified by civilian and military research as useful in predicting aviator performance, including psychomotor coordination, background information, information processing ability, general cognitive ability, and personality traits. Twelve abilities were emphasized by Lochner and Nienhaus (2016) for the job of a pilot, including general cognitive abilities, such as inductive reasoning and deductive reasoning, as well as more specific abilities, such as spatial orientation and control precision.

A noteworthy remark from these reviews is that they considered a broad range of performance determinants, such as cognitive abilities, psychomotor ability, personality, domain-specific knowledge, biographical information, and prior experience. Widening the space of investigations to include a variety of potential factors is advantageous for thoroughly understanding constructs influencing pilot performance, from a multidimensional perspective. Even a small effect or little incremental validity provided by some of these cognitive and non-cognitive constructs can have substantial practical consequences (e.g., Noftle & Robins, 2007; Ozer & Benet-Martinez, 2006).

2.2.3 Modeling Pilot Performance

Pilot performance research is essential to ensure operational performance readiness. Because each job has specific attributes different from other jobs, it may also require distinct methods for assessing its performance 'quality' (i.e., differential job analyses). Perhaps the most valid criterion of pilot performance is the successful and safe accomplishment of the operational mission. However, the measurement of pilot performance is complex and so not clear-cut. As nicely stated by Bates et al. (1997), "No single construct, or operationalization of variables, fully addresses pilot performance. Rather, a multi-disciplinary and multi-modal approach, using significant developments from recent studies, holds the most promise." To conceptualize pilot performance, they suggested a holistic approach that integrates different lines of research in pilot performance, which include pilot training criterion measures, safety studies, pilot perceptions of performance, cockpit resource management, and human factors. Even the performance of pilot applicants during the selection process has caught researchers' attention, and was used as a criterion such that success at final stages is subject to prediction by performance scores at early stages of selection (e.g., Hoermann & Damos, 2019; Hoermann & Goerke, 2014). This is perhaps due to the high cost commonly associated with some assessment methods (e.g., flight simulator, structured interviews led by experienced pilots or psychologists). In view of the nature of the data utilized in the current thesis, training measures of performance were the focus in these investigations. Examples of such measures include pilot performance in the academic phase of training, or the daily flight ratings given to pilot students. Passing or failing the training program and the class rank of students are popular indexes for the training performance of ab-initio pilots.

2.2.4 Predictors of Pilot Performance: Criterion-related Validity

Criterion-related validity is essential for job selection assessment tools. Predictive validity involves assessing how well test scores predict future performance outcomes. Carrying out a local validation study is a constructive way to determine the effectiveness of job selection predictors within an organization under investigation. Oftentimes, however, such a systematic study may not be possible for an organization, possibly due to the inadequacy of needed samples, absence of qualified researchers, or time constraints. Hence, other workable options that may suffice include relying on results from local studies conducted for other flying organizations or, more preferably, results derived from meta-analysis investigations. Pilot performance relations with numerous constructs have been investigated both locally at organizations' empirical examinations level or more broadly at the comprehensive review level. This part will focus on local studies performed at specific settings, and the next section will present a summary of cross-setting meta-analyses.

In addition to psychomotor and cognitive ability tests, many selection methods have been assessed for their effectiveness as predictors for pilot performance. Evidence for predictive validity has been documented for a number of these methods such as assessment center testing, situational judgment tests, crew resource management procedures, work samples, and structured interviews. Moreover, personality traits and biographical data are often used in selection procedure and have been found to have some utility as predictors for future job performance. Hence, I present below a few examples of validation studies assessing the role of some selection methods and personality constructs for predictiveness of pilot performance, followed next by studies assessing psychomotor ability and cognitive ability constructs.

2.2.4.1 Selection Methods

Selection of pilot applicants is typically carried out in a multiphase process. Most often, low-cost methods are incorporated in the early stages of selection, while those associated with higher costs are reserved for a later stage of selection. As an illustration, Hoermann and Goerke (2014) showed that pilot selection in the German Aerospace Center (DLR) follows a multistage procedure, with five main stages: (a) basic pilot aptitude tests, (b) psychomotor tests, (c) an assessment center, (d) a fixed-base simulator, and (e) an interview. The overall selection ratio resulting from this process was estimated to be 11.4% of the total applicants, with average pass rates of 35.7% (basic aptitudes), 87.2% (psychomotor tests), 64.9% assessment center, 77.1% (simulator), and 73.2% (interview). Paullin et al. (2006) recommended a two-stage testing process in the selection of the US Army aviator selection, one for administering measures of cognitive abilities and personality/motivational traits, and the second for performance-based testing of psychomotor and information processing skills.

Assessment center testing has become an essential element of occupational selection. Because this method gives a direct assessment of teamwork-related behavior, its use in aviation and pilot selection is often suggested (e.g., Bartram & Baxter, 1996; Damitz et al., 2003; Hoermann & Goerke, 2014). Damitz et al.'s (2003) study investigated the validity of a newly-introduced assessment center approach for pilot selection in the DLR using 1,036 pilot applicants' sample. A principal components analysis revealed a two component solution that explained 68% of the variance in selection test scores: performance competence and interpersonal competence. The test battery's criterion-related validity was assessed by following up with a subsample of successful applicants using peer pilot and psychologist ratings as criterion measures (N = 73). Results from two performance criteria, as rated by pilot peers and psychologists, respectively, revealed observed correlations of .24 and .27 for the Performance Competence factor, .24 and .09 for the Interpersonal Competence factor, and .29 and .21 for the overall rating derived from the assessment center.

Another method that gained attention in selection research is the *situational judgment test (SJT)*. It differs from traditional cognitive testing in that it contains job-relevant scenarios describing complex situations in the subject occupation. Recently, for example, the USAF added an SJT to the AFOQT Form T, aiming to improve assessment of officership (Carretta et al., 2016). Hunter (2003) developed a 39-item SJT for general aviation pilots. Each item presents an in-flight situation that requires a decision to be taken by the pilot from four alternative solutions provided to the situation. After the evidence of reliability and validity were established in a validation study, the test was further examined for criterion-related validity against the number of accidents and other hazardous aviation events experienced by each participant (N = 115). Results showed that pilots with higher (better) scores on the SJT tended to experience fewer hazardous events (–.215, p = .021), suggesting acceptable validity of the test for predicting pilot performance. Hunter's (2003) SJT was recently shortened to 16 items by Shi (2012), and also yielded acceptable evidence of validity and reliability.

A skill related to situational judgment is what is known in aviation as *crew resource management (CRM)*. CRM describes the efficient use of available resources such as equipment, people, or information to accomplish a flight operation successfully and safely (Driskell & Adams, 1992). Hedge et al. (2000) developed and validated a test of CRM including attributes that are considered crucial for a pilot job. The test was designed with 60 items simulating realistic, albeit difficult, aircrew situations, and representing six main dimensions: problem-solving, decision making, knowledge of how to respond to challenging situations, communication, aircrew management, and interpersonal effectiveness. Due to the length of the test, a short version containing 13 items was also proposed and suggested for operational use. The test's validation study involved a sample of 115 experienced C-130 aircrews from eight Air National Guard bases in the US. Respondents were asked to judge and select the most and least effective response to each situation from five possible answers. The criterion used in the study was a performance measure developed explicitly for the study to assess the performance of aircraft commanders on constructs equivalent to those used in CRM tests they had previously taken (e.g., crew coordination, communication, teamwork, and problem-solving). The commanders were assessed by seven crew members (copilots, navigators, flight engineers, loadmasters, radio operators, and other aircraft commanders) who experienced the flying skills of each aircraft commander and were in a position to assess his performance. When the scores of the CRM tests and the overall ratings of crewmembers were correlated, results yielded observed correlation of .19 for the operational test (13 items) and .14 for the original test (60 items), suggesting an acceptable criterion-related validity of the proposed CRM test.

Across occupations, *work sample* have shown to be among the best predictors of future training and job performance (Schmidt & Hunter, 1998). A similar pattern is frequently observed in aviation, where flight simulator testing administered during the selection process has been found to be a highly valid predictor for flight training performance. For illustration, Woycheshin (2002) examined the validity of flight simulator performance for predicting subsequent training performance. He related scores obtained from different sessions of flying tests via the Canadian

Automated Pilot Selection System (CAPSS), a computerized simulator of a single-engine light aircraft, during the selection process with other scores obtained for pilot students in their primary phase of flying training. Correlations with criteria of pass/fail (.27), course grades (.32), flying performance ratings (.35), and ground training academic averages (.27) had moderate magnitude, although this pattern of associations was still higher than noted for other psychological constructs. In another study, Johnston and Catano (2013) also assessed the CAPSS, along with other predictors, and they found it the best predictor in the study, with adequate power to predict academic and actual flying performance at both early phase of training (r = .40 and .36, respectively) and more advanced levels of training (r = .33 and .18, respectively). The recent ALMamari and Traynor (2019) meta-analysis estimated an observed mean validity of .34 for work sample scores against overall pilot performance criterion. Similar estimates were also revealed in Hunter and Burke's (1994) and Martinussen's (1996) meta-analyses.

One well-known method in selection is the *structured interview*, although the validation of this method in aviation seems to be limited. In a study of USAF pilot applicants, Walters et al. (1993) assessed the incremental validity of a structured selection interview for pilot performance (N = 223). Seven dimension ratings were derived from the interview, tapping educational background, self-confidence, leadership, flying motivation, success in training, and success in flying various classes of aircraft. Results from regression analysis showed that the seven dimensions of the structured interview were predictive of pilot performance (R = .21). However, when the interview's scores were added to a model already containing scores from the AFOQT and a test battery measuring information processing and personality, the interview ratings did not show significant incremental validity.

2.2.4.2 Personality Traits and Biographical Data

Personality assessment becomes a major part in selection in many organizations. Several studies exist attempting to link personality traits with pilots' flight performance. The meta-analysis of Campbell et al. (2009) assessed three constructs of personality, Neuroticism (k = 7), Extroversion (k = 8), and Anxiety (k = 4). Results from the random-effects model suggested observed effects on flight performance of –.15 for Neuroticism, .13 for Extroversion, and –.11 for Anxiety. The recent study of Barron et al. (2016) assessed whether the "big five" factors of personality (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness) and a Machiavellianism trait had any predictive utility for pilot and navigator job performance. Two criteria were used in the study, both of which were extracted from the personal record of each participant, in their Officer Performance Reports (OPR) (presence/absence of a stratification statement and strength of stratification statements). Stratification statement is a numerical value given to the Officers on their OPRs by senior raters or their direct supervisors as an index of overall performance. Concerning pilots, significant relations were found between Agreeableness, Emotional Stability and Extraversion, and strength of stratification statements criteria (observed r = .20–.30), but none was found with presence/absence of a stratification statement.

In another study, Carretta et al. (2014) also assessed the validity of the big five personality factors (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness) for predicting USAF pilot performance by location and three training tracks (primary aircraft training [T-6], the fighter/bomber advanced training track [T-38], the airlift/tanker advanced training track [T-1]) for a 14-year period. Of the 15 correlations in the overall sample (N = 1,524 - 9,396), many were found to be statistically significant, although small in magnitude, ranging from |.01| to |.11|. Openness to Experience was the best predictor for T-6 and T-38 (r = -.06 to -.07), while Conscientiousness was the best predictor of T-1 (r = .11). Results also determined an adequate level of consistency over time for the relations of personality traits to pilot training performance. Last, Hoermann and Goerke (2014) assessed whether scores of pilot applicants in social competence and personality (as measured by the Social Skills Inventory) at an early stage of selection could be as valid as assessment center scores in predicting applicants' final performance ratings in the selection process, and success in pilot training (pass/fail criterion). Results showed a limited utility of social competence scores to predict either applicants' overall performance in the pilot selection process or subsequent training outcome. Nevertheless, the authors concluded that social competence and personality questionnaires could be added as an additional component in the preselection, although they cannot replace the assessment center. Overall, some personality traits have shown significance in predicting pilot performance; however, their observed correlations with criteria are generally small and inconsistent across studies.

It is common in occupation selection to collect biographical data about applicants in order to obtain more background information supplementing other selection criteria. Interestingly, ratings given to some domains of background are found to relate significantly with scores of pilot performance. A series of pilot source studies (Smith et al. 2010; Smith et al., 2013; Smith et al., 2016) exemplifies a direction of pilot research emphasizing the role of pilots' college and piloting backgrounds as influential determinants of future performance. Stricker (2005) assessed the relations of the Biographical Inventory used in the selection of U.S Naval aviation training students with student retention criteria, after factor-analyzing its underlying constructs. The EFA revealed that the inventory featured five factors: commissioned officer, science and engineering interests, flight experience, masculine activities, school athletics. The best two predictors against the two criteria in the study, overall ratings for performance in Naval aviation schools command and primary flight training, were commissioned officer factors (r = .28 and .21) and school athletics factors (r = .14 and .11), suggesting that these two factors account for most of the Inventory's validity. Surprisingly, when the Biographical Inventory assessed as a composite containing all factors, its predictiveness decreased to .03 and .06 for the Naval aviation schools command and primary flight training criteria, respectively.

Of the biographical data, *previous flying experience*, in particular, is shown to be an influential variable for predicting pilot performance. Due to this established conclusion, the USAF has included a score of previous flying experience in the Pilot Candidate Selection Method (PCSM) (Carretta, 2011), along with the Pilot composite of the AFOQT and a psychomotor-perceptual composite from the Test of Basic Aviation Skills (TBAS). As an illustration, Johnston and Catano (2013) examined the validity of three classes of predictors, previous flying experience, simulation performance (CAPSS), and cognitive ability, for two academic and two flying performances of military pilots at their entry-level and intermediate-level of training (N = 150-300). Previous flying experience predicted success in early academic and flying performance (r = .40 and .35) but not in the more advanced levels of pilot training.

2.2.4.3 Psychomotor Ability

The psychomotor ability construct has received considerable research effort in aviation, and has been recognized as a critical ability for pilot performance (e.g., Fleishman, 1956; Griffin & Koonce, 1996; Wheeler & Ree, 1997). Most pilot performance-based selection batteries include tests assessing this ability. By way of example, the USAF Basic Attributes Test (BAT; Carretta et al., 2000) and its successor the Test of Basic Aviation Skills (TBAS; Carretta, 2005), as well as the U.S. Navy's performance-based ASTB (Phillips et al., 2011) are all test batteries highly loaded with psychomotor-perceptual abilities. Three studies of these instruments are described below. Nye et al. (2018) assessed the possibility of obtaining incremental validity beyond a general ability score when using psychomotor ability scores derived from the performance-based battery of the U.S. Navy's ASTB, proposing that these scores should be more correspondent than *g* to tasks performed on the flying job. Five scores representing psychomotor ability (Directional Orientation Test [DOT] Total, DOT Time, Airplane Tracking Test [ATT] Composite, Vertical Tracking Test [VTT] Composite, Emergency Scenario Test [EST] Total) and a score of *g*, as indexed by the Academic Qualification Rating (AQR) of the ASTB, were correlated with five performance criteria collected during flight training: Contact Stage, Instruments Stage, Navigation Stage, Formation Stage, Navy Standard Score. Both classes of predictors, *g* and psychomotor tests, predicted the criteria almost comparably. Correlations of the AQR score with criteria ranged from .04 (Navigation stage) to .34 (Instruments stage and Navy Standard Score), while performance-based score correlations varied from .03 (DOT total/Navigation stage) to |.31| (ATT composite/Instruments stage) across performance criteria. Regression-based analysis showed that psychomotor ability scores added significant incremental validity to a measure of *g*, with increases in R^2 ranging from .08 to .10 across criteria.

Caponecchia et al. (2018) examined the ability of the WOMBAT pilot selection test to predict performance outcomes in a sample of ab-initio pilots at the University of New South Wales (UNSW) Flight Operations Unit ($N = 60$). Correlating three scores from WOMBAT tests, Tracking Score, Total Bonus Score, Overall Score, with 15 instructor ratings at RPL level yielded average coefficients of .21, .28, and .27, respectively. Similarly, correlating the same three scores with 12 instructor ratings at PPL level yielded average coefficients of .20, .38, .35, respectively. Despite the seeming higher correlations at the PPL level than the RPL level, the number of statistically significant results at the RPL level was many more than at the PPL level (13 vs. 4). Hoermann and Guan (2002) assessed the validation of an ab-initio pilot selection method after transferring it from

Germany to China. The adaption involved many validation processes, among which was following up with two samples of student pilots ($N1 = 125$; $N2 = 200$) to evaluate the criterion-related validity of different measures in the test batteries, including aviation-related knowledge, operational abilities, personality, and psychomotor abilities. Criteria in the study included scores representing the Flight Screening phase, Primary Training phase, and Advanced Training phase. Results showed that apparatus-based tests of psychomotor coordination and multiple task performance provide the best predictions among the examined predictors, with correlation coefficients across two samples, and three criteria ranging from .19 to .32 for psychomotor coordination and from .23 to .49 for multiple tasks performance. These few examples from different flying organizations indicate that psychomotor ability constructs are significant contributors for pilot performance and their validity magnitudes tend to be higher than many of the predictors presented in this section.

2.2.4.4 Cognitive Abilities

Zierke (2014) assessed the predictive validity of the DLR assessment using a sample of 402 students, with particular attention to subject-specific knowledge tests. The training outcome of pass/fail was used as a criterion in the study. With the exception of English test scores and English school grades, the remaining three knowledge tests (Mathematical Reasoning, Technical Comprehension, and Physics and Technical Basics) and two school grades (Mathematics and Physics) showed significant relations with pilot performance (observed $r = .09$ to $.14$). In contrast, for cognitive ability tests, except for significant relations of Point Position and Memory Search tests with the training outcome (observed $r = .12$ and $.08$; respectively), relations of the remaining five tests (Visual Perception, Symbol Concentration, Running Memory Span, Mental Arithmetic, Dice Rotation) with the outcome were not significant. After correcting the correlations for range-restriction and dichotomization, however, the multiple correlations of a model containing all predictors (seven ability tests and four knowledge tests) reached .55.

In a sample of 108 student pilots from the South African Air Force (SAAF), Kock and Schlechter (2009) examined the extent that fluid intelligence and spatial reasoning are predictive of flight performance criteria. Including both predictors in a multiple regression equation yielded multiple correlations of .35, .20, and .23 for the three criteria of practical flight training, ground school training, and officers' formative training, respectively. Also, the spatial ability was found to increment validity beyond that of fluid intelligence for predicting flight training performance.

Aiming at improving the effectiveness of pilot selection in the China Air Force, Wang et al. (2018) assessed the possibility of predicting pilot students' success in training (i.e., pass/fail) with tests of spatial working memory (WM) and visual perspective taking (VPT). Results indicated significant positive correlations between working memory and success in both primary (r = .15) and advanced (r = .18) phases of training, confirming the predictive utility of the working memory construct. For VPT, no significant relations were found with either of the two performance criteria.

King et al. (2013) investigated whether two standard cognitive psychological test batteries (the Multidimensional Aptitude Battery [MAB-II] and the MicroCog) have noteworthy predictive power for initial pilot training performance (N = 5,582-12,924). They associated cognitive scores with seven performance criteria, including three training completion criteria (all eliminees, eliminees with training deficiency, and eliminees dropping on their own request), academic grades, daily flying grades, check ride grades, and class rank. The relations of cognitive variables with performance criteria indicated small but statistically significant relationships. For the MAB-II, across the seven criteria proposed for the study, most observed validities of full-scale IQ (.03 to .23) were slightly higher than those of performance IQ (.04 to .16), and both were mostly higher than those of verbal IQ (.02 to .22). For the MicroCog battery, the five content-specific scores (attention/mental control, memory, spatial processing, reasoning/calculation, reaction time), as well as the four broad composite scores (information processing speed, information processing accuracy, general cognitive functioning, general cognitive proficiency) all correlated significantly with the five performance criteria (r = .02 to .11), excluding five non-significant relations with students dropping out on request (r = .00 to .02). In general, the relative importance of predictors tended to differ as a function of the performance criterion used in the validation procedure.

In another validation study of the MAB–II, Carretta et al. (2014) assessed the validity of its three summary scores (verbal IQ, performance IQ, and full-scale IQ), along with personality scores, for predicting pilot performance in a sample of 9,641 USAF student pilots assessed between 1995 and 2008. Across three performance criteria extracted from three training tracks (primary aircraft training [T-6], the fighter/bomber advanced training track [T-38], the airlift/tanker advanced training track [T-1]), the MAB–II composite scores showed significant relations with all three training tracks, with small differences in the magnitude of validities (observed r = .06 to .14). The magnitudes of the validities were generally higher for cognitive constructs than for personality traits, although all were rather small even after correction. Overall, results showed consistency over time for the relations of cognitive ability and personality traits to pilot training performance.

In addition, other related cognitive constructs were also examined for possible predictive relations with flight performance measures such multitasking ability and executive function. Barron and Rose (2017) compared the validity of scores from a multitasking assessment (math, memorization, and monitoring tasks presented concurrently) and the individual single-tasks that formed the multitasking assessment for two criteria of academic and flying performance. Results showed that scores of initial and post-practice concurrent multitasking predicted both flying (*r* = .19 and .23) and academic (*r* = .12 and .21) performance, while scores of serial single-tasks and their combination predicted only academic performance (*r* = .04 to .16). Causse et al. (2011) examined the concurrent validity of three low-level executive functions (updating in working memory, inhibition, set shifting), as well as some other cognitive and background variables, for flight performance (*N* = 24). The criterion used in the study was pilots' performance on a PC-based flight simulator, which involved the ability of pilots to fly without substantial deviation from the path given in the mission plan. The more deviation from the path, the lower the score pilot was given. Among the three executive functions, only updating in working memory was significant in predicting pilot performance, in addition to reasoning and previous pilot experience, while inhibition and set shifting did not show valid predictiveness. The authors reasoned that those predictors' lack of validity could be due to the characteristics of the flight scenario used in the study, which may have required more updating ability and less inhibition and set shifting abilities.

2.2.4.5 Concluding Summary

The examples presented above show a diverse range of cognitive and non-cognitive constructs that have been linked to flight performance in an attempt to understand how they influence pilot performance. Similarly, a wide array of performance measures has been used in studies, which indicates a thoughtful attempt to widen the investigated criterion space of pilot performance. Examples for criteria utilized in the presented studies include training outcome (pass/fail), academic grades, daily flying grades, advanced training track, psychologist ratings, and the number of accidents or other hazardous aviation events experienced by a pilot. This can be essential in the design of validation studies as the validity of each predictor-criterion combination tends to vary by the specific criterion measure used and hence, different inferences may be derived.

It is important to understand which variable is best at predicting the specific criterion. For example, the U.S Navy uses two different composite scores derived from the ASTB to predict success in ground school and primary flight school. The Academic Qualifications Rating (AQR) composite is used to predict performance in ground school, while the Pilot Flight Aptitude Rating (PFAR) composite is used to predict success in primary flight school (Paullin et al, 2006). Researchers in pilot performance manipulate a variety of predictor-criterion combinations and are aware of psychometric aspects of criterion-related validity. The results accumulated from these efforts are expected to lead to an improvement in selection models and enhance the understanding of human factors surrounding pilot performance. Additionally, although the studies presented above are not meant to be an exhaustive list, they clearly show that research effort in the field is carried out in many places and organizations (e.g., the U.S., Germany, South Africa, China). This demonstrates that the interest in aviation psychology is worldwide and not limited to a certain country or organization.

Of note, although cognitive constructs are seen among the best predictors of pilot performance, the correlation coefficients in the validation studies are mostly small. The majority of best cognitive predictors come between the value of .10 and .20, and very few exceeded this range, although many were below it. According to Cohen's (1988) benchmarks for the interpretation of the effect size of correlations (.10, .30, and .50 indicate small, medium, and large, respectively), these are considered small correlations. However, Gignac and Szodorai's (2016) recent study noted that this guidelines in applied practice is arguable and rarely achieved, and suggested instead a more practical values (.10, .20, and .30 indicate relatively small, typical, and relatively large, respectively), after assessing a large sample of previously published meta-analytically derived correlations (N = 708). Based on this suggested guideline, many of the effect sizes reviewed in this section can be seen as typical or moderate. However, I argue that these values might be still exigent in pilot performance studies. Future research may focus in the criterion-related studies in aviation to assess whether a less exigent guideline can be more representative to the effect sizes from a normative perspective.

2.2.5 Predictors of Pilot Performance: Meta-analyses

Meta-analysis is a useful technique that aggregates studies from different settings, samples, and years to obtain summary statistics for the relations between variables with evidence of validity generalization. In order to demonstrate the importance of cognitive abilities for flight performance, I present examples from previous meta-analysis studies. In aviation, there exist several meta-analyses that assessed the predictive validity of ability tests for criteria measuring performance in flying training programs and actual jobs. Hunter and Burke (1994), Martinussen (1996), and ALMamari and Traynor (2019) are examples for meta-analysis at construct and/or composite level, while Martinussen and Torjussen (1998), Burke, Hobson, and Linsky (1997), and ALMamari and Traynor (2020) are examples for meta-analysis at individual test level. Damos (1993) and Campbell et al. (2009) presented examples of meta-analysis for specific constructs: multitasking and personality, respectively. In addition, some meta-analyses were concerned with specific areas within aviation such as the effectiveness of flight simulator (De Winter et al., 2012; Hays et al., 1992; Vaden & Hall, 2005) and crew resource management (CRM) (O'Connor et al., 2008) selection procedures.

Table 1 summarizes the results of six meta-analyses (ALMamari & Traynor, 2019; 2020; Hunter & Burke, 1994; Lynch, 1991; Martinussen, 1996; Martinussen & Torjussen, 1998) that examined cognitive ability constructs, which is the focus of this thesis. More than 73 cognitive (and some non-cognitive) variables were incorporated across the six meta-analyses, which indicates the effort being made to thoroughly understand cognitive abilities' relations with flying skills. Table 1 groups these variables according to the validity magnitude. It is noted that the observed mean correlations of ability-performance relationships rarely exceed .30 value, with only five cases exceeded this value, representing work sample (two studies), gross dexterity (two studies), and a combined index of cognitive and psychomotor composites. The second class of best predictors ranged between .25 to .29, consisting of constructs of mechanical ability, reaction time, biodata inventory (one of four studies), instrument comprehension, general information, training experience. The .20 to .22 range of mean correlations included constructs of aviation information (three of four studies), combined cognitive tests, biographical inventory (one of two studies), perceptual speed, psychomotor/information processing.

Table 1. *Summary of Six Meta-analyses on the Predictive Validity of Selection Measures for Pilot Performance*

Mean *r*	**Predictors**
.31-.34	Job Sample; Work Sample; Gross Dexterity; Combined Index
.25-.29	Mechanical Ability; Reaction Time; Biodata Inventory; Instrument Comprehension; General Information, Training Experience
.20-.22	Aviation Information; Combined Cognitive Tests; Aviation Information; Biographical Inventory; Aviation Information; Perceptual Speed; Psychomotor/Information Processing
.15-.19	Spatial Ability; Mechanical ability; Acquired Knowledge; General Ability; Motor Abilities; Scale Reading, Instrument Comprehension; AFOQT Pilot Composite; English test; ASTB Pilot Composite; Academics average; AFOQT Pilot Composite
.10-.14	Ravens Matrices; Perceptual Processing; Data Interpretation; General Ability; Intelligence; Personality; Table Reading; Aviation Information; Verbal Ability; English test, DMT NPI [personality test]; Arithmetic Reasoning; Quantitative Ability; Mechanical Comprehension; Fine Dexterity; Personality; Spatial Orientation; Mirror Tracing; Controlled Attention; Rotated Blocks
.05-.09	Number Series; Reversal Test; Math Knowledge; Block Counting; Electrical Maze; Hidden Figures; Education average; Sorting Test; DMT 10 [personality test]; Mathematics test; Reading Comprehension, General Science
.00-.03	Verbal Analogies; Block Counting; Figure Pattern; Word Knowledge
< .00	Paper Forming; Rotating Patterns; Numbers test; Age

By contrast, the weakest predictors among cognitive variables included on this summary are four AFOQT subtests (Verbal Analogies, Block Counting, Figure Pattern, Word Knowledge) with validity estimates lower than .03, as well as three subtests from the Norway Air Force pilots' selection battery (Paper Forming, Rotating Patterns, Numbers) that even had, unexpectedly, negative correlations with flight performance criteria. Of note, composite scores that typically contain a mix of cognitive constructs and arguably, are more powerful predictors did not predict better than scores from some of the best individual ability tests.

2.3 The Air Force Officer Qualifying Test (AFOQT)

2.3.1 Overview of the AFOQT Structure

The AFOQT is a multiple-aptitude test battery composed of 10 ability subtests in the most recent version, Form T. The battery is designed to assess a variety of cognitive (e.g., *g*, verbal, quantitative, spatial) and aviation-related knowledge (e.g., Aviation Information) constructs. The AFOQT is used for officership qualification and initial job placement for officers selected in the U.S. Air Force. The subtest configuration of AFOQT has changed from 16 ability subtests (Form O, P, Q), to 11 subtests (Form S), and most recently, to 10 (Form T). Its reliability has been studied extensively (e.g., Berger et al., 1990; Carretta et al., 2016; Glomb & Earls, 1997; Skinner & Ree, 1987) ranging from .730 (RC) to .913 (IC) in the most recent form (Carretta et al., 2016), and it has been validated for officer training performance (Roberts & Skinner, 1996) and several Air Force occupations (e.g., Arth, 1986; Carretta, 2010; Finegold & Rogers, 1985). Most of the AFOQT validation efforts were directed to aviation jobs such as flying, navigation, and air battle management (e.g., Barron et al., 2016; Carretta, 2008; Carretta & Ree, 1995; Johnson et al., 2017; Olea & Ree, 1994). For job decision purposes, operational composites are typically derived using different sets of AFOQT subtests; currently, seven overlapping composite scores are computed from AFOQT Form T: Pilot, Combat Systems Officer, Air Battle Manager, Academic Aptitude, Verbal, and Quantitative (Aguilar, 2017). In addition to cognitive ability tests, Forms S and T also included the *Self-Description Inventory+*, a personality inventory measuring the big five factors (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness) and Machiavellianism trait (Aguilar, 2017). Moreover, a *Situational Judgment Test* was also introduced in Form T as an experimental assessment, which reflects a tendency to make use of different methods for enhancing the dependability of selection assessment (Carretta et al., 2016).

2.3.2 AFOQT Across Six Forms

The AFOQT configuration of 16 subtests appeared first in Form O (Skinner & Ree, 1987) and was maintained across two succeeding forms; Form P (Berger et al., 1990) and Form Q (Glomb & Earls, 1997). The subsequent versions of AFOQT excluded some of the subtests; five subtests were omitted from Form S (Drasgow et al., 2010), and six subtests were eliminated from the current Form T (Barron et al., 2016). Table 2 on the next page shows the subtest composition of AFOQT across the three versions. Appendix A presents examples of test items typically found in the AFOQT. Some subtests in the long AFOQT version were carried forward from earlier versions (Skinner & Ree, 1987), and the retained subtests in later versions were all taken forward from the extended version. Because of this, accumulating AFOQT intercorrelation matrices from different versions for the intended meta-analytic SEM study in this thesis can be justified, especially with the rigorous equating and alignment procedure regularly conducted when replacing an old version with a newer version (e.g., Glomb & Earls, 1997).

Table 2. *The AFOQT Configuration Across Different Forms and Subtests CHC-based Classification, Grouped by the Five-factor Model*

Subtest		Q	S	T	CHC-Based Classification	Construct
Verbal Analogies	VA	X	X	X	Fluid--Induction	Measures the ability to reason and recognize relationships between words.
Reading Comprehension	RC	X	_	X	Acquired Knowledge--Verbal Ability--- Reading and Writing----Reading Comprehension	Measures the ability to read and comprehend paragraphs.
Word Knowledge	WK	X	X	X	Acquired Knowledge--Verbal Ability---Comprehension Knowledge----Lexical Knowledge	Measure the ability to understand written language through the use of synonyms.
Arithmetic Reasoning	AR	X	X	X	Fluid--Quantitative Reasoning	Measures the ability to understand arithmetic relationships expressed as word problems.
Data Interpretation	DI	X	_	_	Fluid--Quantitative Reasoning	Measures the ability to interpret data from graphs and charts.
Math Knowledge	MK	X	X	X	Acquired Knowledge--Quantitative Ability---Mathematics Knowledge	Measures the ability to use mathematical terms, formulas, and relationships.
Mechanical Comprehension	MC	X	_	_	Acquired Knowledge--Domain-Specific Knowledge---Sciences----Mechanical Knowledge	Measures mechanical knowledge and understanding of mechanical functions.
Electrical Maze	EM	X	_	_	Visual Processing--Spatial Scanning	Measures spatial ability based on choice of a path through a maze.
Block Counting	BC	X	X	X	Visual Processing--Visualization	Measures spatial ability through analysis of three-dimensional representations of a set of blocks.
Rotated Blocks	RB	X	X	_	Visual Processing--Visualization	Measures spatial aptitude by requiring mental rotation and manipulation of objects.
Hidden Figures	HF	X	X	_	Visual Processing--Flexibility of Closure	Measures spatial ability by requiring the detection of simple figures embedded in complex drawings.
Scale Reading	SR	X	_	_	General Mental Ability	Measures the ability to read scales and dials.
Table Reading	TR	X	X	X	Processing Speed--Perceptual Speed--- Scanning	Measures the ability to extract information from tables quickly and accurately.
Instrument Comprehension	IC	X	X	X	compound-Acquired Knowledge & Visual Processing	Measures the ability to determine aircraft attitude from illustrations of flight instruments.
Aviation Information	AI	X	X	X	Acquired Knowledge--Domain-Specific Knowledge---Sciences----Physical Sciences Knowledge (Applied)	Measures knowledge of general aviation concepts and terminology.
General Science*	GS	X	X	X	Acquired Knowledge--Domain-Specific Knowledge---Sciences----General Science Knowledge	Measures knowledge and understanding of scientific terms, concepts, principles, and instruments.

Note. Description of the subtests was obtained from Carretta and Ree (1995); the CHC-based classification of the subtests was obtained from Stanek and Ones (2018) and KC. Stanek (personal communication, March 09, 2020)

2.3.3 AFOQT Factor Structure

Factor analytic studies of AFOQT are relatively small in number. There exist five factor-analytic investigations for the AFOQT, of which only one study assessed the data by EFA, while the rest applied a confirmatory approach through CFA procedures. This group of studies represents the three AFOQT versions described above: the 16-subtest version, Form S, and Form T. Skinner and Ree's (1987) is the only study that assessed AFOQT factor structure (Form O) through an exploratory technique, and apparently, is the reference most cited to justify the factor structure choice in later studies. In their study, Skinner and Ree (1987) explored one to six factors solution for AFOQT data via principal factors analysis and oblique rotation (N = 3000), and the results supported a five-factor structure to be the best representation of the data. The five dimensions of broad ability were named Verbal, Quantitative, Space Perception (spatial and mechanical), Perceptual Speed, and Aircrew Interest/Aptitude. Intercorrelations among the five factors were not as strong as might be expected with range from .22 (Verbal and Perceptual Speed) to .50 (Space Perception and Perceptual Speed), and an average of .36, a pattern that was interpreted as a possible existence of two to three higher-order factors if it was taken further for second-order factor analysis (see also Warne & Burningham, 2019).

The same data were later assessed by Earles and Ree (1991) for the purpose of estimating the general factor of ability (or *g*) and whether different estimation methods (i.e., unrotated principal components, unrotated principal factors, and variants of hierarchical factor analysis) would yield different estimates of *g*. Results showed that all methods produced a somewhat similar estimation of *g*, with correlations ranging from .980 to .999. Carretta and Ree (1996) further examined the same data of Form O (N = 3000) using CFA procedures. Seven CFA models were assessed in the study including, the single-factor, four-factor model (based on operational composites), orthogonal five-factor (based on previous EFA study), bi-factor (three different models), and higher-order factor. The best-fitting model among the seven was a bifactor model involving the five group factors suggested by Skinner and Ree (1987), resulting in CFI of .96, RMSEA of .07, and SRMR of .03. However, the other two bifactor models, as well as the higher-factor model (Vernon-like model) showed acceptable fit for the data (CFI > .95 and RMSEA < .08).

Drasgow et al. (2010) examined the factor structure of AFOQT Form S data (11 subtests) that was collected from 12,511 USAF officer applicants, with additional goals of establishing measurement invariance across gender (male/female) and race (White/African American/Hispanic/Other groups). A total of nine models were specified in the study, seven of which were defined using item parceling procedures. Similar to Carretta and Ree (1996), results indicated that the data were best represented by a bifactor model containing a general intelligence factor and five content-specific factors representing verbal, quantitative, spatial, perceptual speed, and aircrew aptitude/interest (RMSEA = .05, CFI = .98, and SRMR = .057). Measurement invariance of the AFOQT across gender and racial/ethnic groups was also established in the study.

In the context of validating AFOQT Form T, Carretta et al. (2016) assessed five different factor structures of AFOQT data: single-factor, four correlated factors, five correlated factors, higher-order four factors, and higher-order five factors. Differing from the other two studies, no bifactor models were tested, possibly due to the insufficient number of subtests variables (10 subtests). This study, akin to Drasgow et al.'s (2010) study, applied the item-parceling technique to provide multiple composites for each subtest to enable factors' specification for CFA. Across two independent samples ($N1$ = 5,681; $N2$ = 5,199), the best-fitting model was a higher-order model with *g* at the top and five lower-order factors at the bottom (verbal, math, spatial, aircrew, and perceptual speed).

In light of the preceding, it is clear that the effort to examine the AFOQT factor structure has continued across the forms. However, there are three limitations to note about previous factor-analytic studies. First, EFA investigations of AFOQT data are rare during the long history of AFOQT use and the regular revisions implemented to the test. Further, a long time has passed since the only available EFA study was reported (more than three decades ago), although it is often used as a theoretical basis for suggesting a five-factor model for AFOQT data. EFA studies are vital, particularly when an instrument undergoes continued revision as in the case of AFOQT. Second, two of the three CFA studies used multi-item composites (i.e., parcels) method to deal with the issue of the small number of indicators available for the AFOQT short versions (Forms S and T). Although item parcels in SEM have been common in research, the use of this procedure is not recommended with multidimensional data (see Bandalos, 2002; Matsunaga, 2008; Sass & Smith, 2006), as is the case with cognitive test batteries. It can be used, however, when data are nonnormally distributed, are coarsely categorized, have a small variable to sample size ratio, or are believed to produce better model fit over solutions at the original item level (Bandalos, 2002). Third, previous CFA studies utilized item-level data for assessing AFOQT factor structure, which was facilitated by parceling. Replicating some of these models with subtest-level scores (obtained from subtests' correlation matrices) instead of item-level data (obtained from the row scores) can provide further support for the AFOQT's factor structure. The use of test batteries' subtest scores to assess a different aspects and frameworks of cognition has been common in recent psychometric studies (e.g., Canivez et al., 2017; Dombrowski et al., 2018; 2019; McGill, 2020). Consequently, the AFOQT is still in need of (1) EFA investigation, (2) CFA investigation without relying on item parceling procedures, and (3) cross-validation of the chosen model(s). These goals can be addressed reasonably via the set of studies proposed for this thesis that include a range of modeling techniques and comparisons. The preliminary meta-analytic SEM study should close the gap noted concerning the lack of a recent EFA study on AFOQT data.

2.3.4 CHC-based Classification of AFOQT Subtests

The recent adaption of the Cattell-Horn-Carroll (CHC) taxonomy to industrial, work, and organizational psychology (Stanek & Ones, 2018) can be useful to categorize AFOQT subtests into broad ability factors. Although the current terminology used in AFOQT studies for describing the ability factors seems revealing, more use of the common language articulated in intelligence and educational research can be even better and would help to harmonize terminologies across domains. Fourteen of the 16 AFOQT subtests are readily-categorized in the Stanek and Ones's (2018) modified taxonomy of the CHC model. The classification of the other two subtests (DI and SR) was completed with generous help offered by K. C. Stanek. Accordingly, a CHC-based categorization is now available for all AFOQT subtests, as presented in Table 2. Of the 16 subtests, seven are, essentially, indicators of Acquired Knowledge (RC, WK, MK, MC, IC, AI, and GS) and five are indicators of Visual Processing (EM, BC, RB, HF, and IC), while the remaining subtests are indicators of Fluid Ability (VA, AR, and DI), Processing Speed (TR), and general mental ability (SR). Although the SR test involves a number of abilities such as visual processing, processing speed-perceptual speed, and fluid-quantitative reasoning, such a heterogeneous mix typically produces a measure of general mental ability, rather than any other specific ability (K.C. Stanek, personal communication, March 09, 2020). This might explain the high criterion-related validity often found for this test against several job criteria.

2.3.5 Pilot Performance Measures Associated with AFOQT

One notable feature in the AFOQT's validation studies is the wide variety of flight performance measures exploited. Diversifying performance criteria in validation studies is an integral component and provides more strength to the validity evidence. AFOQT researchers have made an effort to test the scores' validity for predicting different flight performance criteria. Perhaps no test battery in aviation has been studied as extensively as AFOQT, with criterion-related validity established through such a broad array of flight performance measures. ALMamari and Traynor (2020) provided a useful summary of the criteria used in previous AFOQT validation efforts. For illustrative purposes, ten criteria are exemplified below, representing different phases of training.

(1) *Training Outcome (pass/fail)*. This is a traditional criterion in aviation and is used heavily in validation studies. It is easy to collect, and provides an index for a trainee's overall performance. However, an issue frequently raised with this criterion is the effect of dichotomization (pass/fail) on correlation coefficients, which attenuate the true relations between variables.

(2) *Academic Grade (Ground School)*. This is also a typical criterion in validation studies. It is easy to collect as compared to more sophisticated performance measures. Correlation coefficients with this kind of measure tend to be higher than with actual flying criteria.

(3) *Performance Composite*. Scores representing a composite comprising multiple sources of performance assessment is another useful index of flying performance. Composite scores may include both actual flying and academic grades. When constructed comprehensively, composite scores can be a better indicator of overall flying performance than some other criteria. If only a certain dimension of performance is considered in construction of the composite, such as academic

performance or daily flight performance, then the interpretation should consider the targeted dimension rather than overall performance.

(4) *Class Rank*. This criterion uses the rank order of student pilots in their classes of the study program, based on their overall performance during training. It can be an alternative to training outcome criterion (pass/fail), overcoming that criterion's limitation due to dichotomization. The class rank order gives an indication of overall performance during training, with all aspects considered in the evaluation (e.g., academic, flying, fitness).

(5) *Daily Flight Rating*. This is an important criterion, relying on performance ratings given to a pilot trainee by his instructor on a daily basis. It gives good profile tracking for the pilot's performance throughout the training period and shows their progress fluctuation. Scores representing an average of daily ratings are often computed at each primary phase of training (e.g., basic, advanced).

(6) *Check Flight Rating*. This is a formal hands-on flying exam, programmed periodically at the end of each primary block of training. The evaluators in this type of flying exam are usually an instructor with a high level of flying experience who is not the daily instructor whom the student is flying with, in order to avoid the influence of daily interaction between the trainee and the instructor.

(7) *Flying Hours*. This criterion is derived by assessing the difference between the number of flying hours taken by a pilot to reach a certain level in flying (e.g., flying solo, finishing navigation tasks, graduating) and the average number of flying hours taken by other pilot students to reach that level. Hence, the pilot's performance is rated by the difference between his actual flying hours and the average flying hours. This method might be useful to estimate cost savings in training as a result of cognitive testing.

(8) *Attrition*. Quitting a pilot training program before completion is a negative outcome that is sometimes used as a performance indicator. While lack of flying ability may be the most frequent reason for attrition, medical lack of fitness, behavior or ethical concerns, and family affairs are some other reasons for attrition (e.g., King et al., 2013; Thomas, 2009).

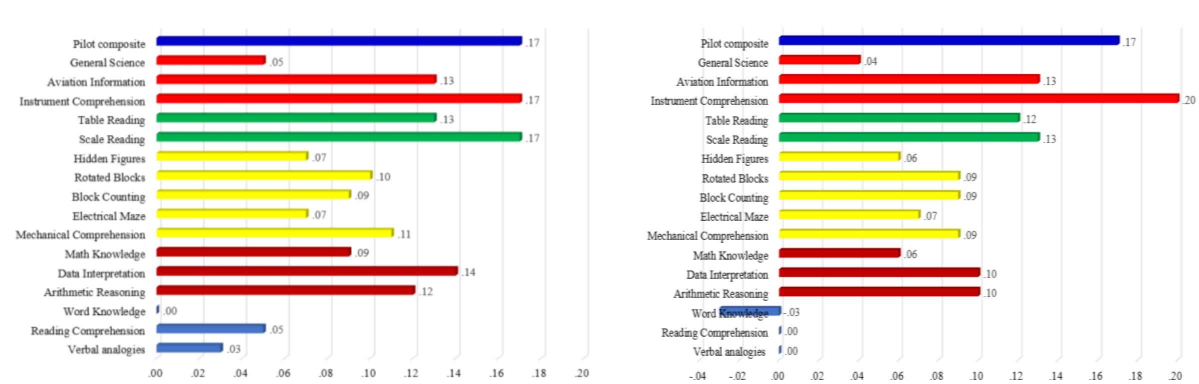
(9) *Advanced Training Recommendation Board (ATRB)*. This is an additional form of performance measure during training and is used as an alternative to the passing/failing criterion. It relies on the assignment granted to pilots after training based on a recommendation from the Advanced Training Recommendation Board (ATRB). The decision about later assignments is usually dependent on a comprehensive assessment of trainees' records during training, and so can be an even better indicator for overall performance than the pass/fail criterion.

(10) *Officer Performance Reports (OPRs)*. This criterion appeared in some recent studies in an attempt to obtain a rating of a pilot's actual job performance. This is a measure of overall performance extracted from the personal pilot record for their flying and non-flying attributes, as rated by supervisors or senior raters.

2.3.6 Predictive Validity of AFOQT Scores: Meta-analysis

Predictive validity investigations are critical and support an instrument's psychometric properties. Numerous validation studies have been published and reported for the AFOQT in the effort of validating its subtests and composites for different criteria of job performance (e.g., Carretta, 2010; Johnson et al., 2017), including flight performance. The AFOQT data utilized in this thesis are examples for such validation efforts for flying job. Due to the variation occasionally found in primary studies' results, Meta-analytic estimates can be more reliable than those obtained from single studies (e.g., Lipsey & Wilson, 2001). A meta-analytic estimate is pooled from several primary studies that are collected through different years of studies, settings, populations/subpopulations, culture, age group, gender, or instrument tools. It allows derivation of one average estimate originated from data of several single studies to investigate a question related to the differences observed between these studies and whether these differences are due to pure chance. Hence, the evidence derived from such a variety of studies for the investigated variables is often seen stronger than that is provided by one single study (Lather, Farkouh, Pogue, & Yusuf, 1997). The recent meta-analysis of ALMamari and Traynor (2020) has synthesized AFOQT literature of pilot performance and provided a quantitative summary for its predictive validity. Therefore, here the presentation focuses on the findings revealed by this meta-analysis, rather than tracing voluminous AFOQT individual studies.

The ALMamari and Traynor meta-analysis assessed AFOQT validation studies that spanned nearly five decades of research. AFOQT scores were examined for their relations with two performance outcomes, an index representing overall flying performance and a Pass/Fail training outcome criterion. Much similarity was noted between the results for the two criteria, particularly as concerned the rank order of subtest validity magnitudes. Based on 32 independent samples from 26 studies, most AFOQT scores' relations (16 subtests and Pilot composite) with pilot performance measures were significant, but with varying degree of predictive power and varying evidence of generalizability across settings. Figure 1 displays the effect sizes of AFOQT subtests (N = 6 to 12) and the Pilot composite (N = 23 to 25) for predicting the two performance criteria.



Adapted from ALMamari & Traynor (2020)

Figure 1. Predictive Validity of the AFOQT Scores for Overall Flight Performance (Left) and Training Outcome (Right)

Overall results showed that subtests indicating aviation-related acquired knowledge and aptitude (AI and IC) and perceptual speed (SR and TR) are the best predictors among the 16 subtests, supported with evidence of generalizable validity. The next-best predictors are two subtests from the quantitative ability domain (AR, DI) that also showed greater validities than the remainder of subtests in the battery. The study emphasized the role of these particular three constructs, and suggested that a selection test battery specifically designed for the flying job should include subtests representing these constructs to further its predictive validity. Subtests demonstrating verbal ability (VA, RC, WK) had the weakest relations with flight performance and showed the lowest generalizability evidence across settings. For spatial ability subtests in the AFOQT (MC, EM, BC, RB, and HF), all five subtests displayed significant validities with evidence of generalizability, and also had the least variability around the means of estimates, in an indication of homogeneity with flight performance measures. However, the observed mean validities of spatial ability subtests were generally lower than those obtained from subtests of other constructs (i.e., perceptual speed), a finding that may require further verification and more empirical studies because it differs from that of other primary studies seemed to establish a more solid role of this construct in aviation job performance (e.g., Barron & Rose; Kock & Schlechter, 2009).

This meta-analysis was informative in synthesizing dozens of AFOQT studies, which enables more certainty about subtests' relations with flight training performance. However, although inferences about the significance of cognitive abilities for pilot performance were possible to a certain degree, there remains a need to capture the latent cognitive traits with more precision. This can be achieved relatively well by using SEM procedures that necessitate conceptualizing each construct in the model with multiple indicators, so its operationalization is complete and representative. Therefore, for a better understanding of the associations between cognitive variables and performance variables, the SEM procedures applied in the four studies of this thesis can be more revealing and demonstrative.

2.3.7 Studies Containing the Data Sets Used in the Current Thesis

The data sets utilized in this thesis were obtained from different sources, all containing correlational data linking AFOQT subtests scores with flying performance scores. Given the well documented validities of AFOQT subtests for pilot performance, the focus, should now be directed to the latent factors underlying these subtests. The previous review of the literature indicates that SEM-based investigations of ability-performance relationships in aviation are lacking. This means that most of our understanding has relied on the observed scores derived from the assessment methods, which provide, given their nature, limited inference about the more abstract level of broad abilities. On these grounds, this thesis presents four SEM-based investigations to fill this gap noted in the literature and enable more sound inference about predictive relations at the construct level. As will shortly be presented, the studies from which correlation matrices were directly reproduced utilized different statistical techniques to associate the variables, which might have been sufficient to achieve those studies' intended functional goals. Yet, in a theoretical sense, there is a need to understand relations between possibly broader, well-defined constructs to assess the extent to which broad abilities influence performance in flying.

The data set utilized in the primary validation study of this thesis was analyzed in two previous studies, Carretta and Ree (1995) and Johnson et al. (2017). While Carretta and Ree (1995) had a broad objective for assessing the predictive validity of all 16 AFOQT subtests for pilot performance, the recent study of Johnson et al. (2017) focused specifically on validating the spatial ability and perceptual speed subtests and examining whether they have any incremental validity over that obtained by other subtests in the battery. The two studies primarily relied on correlations and/or regression procedures for analysis, which were adequate for the validation goals planned by the studies.

For the cross-validation study, three datasets (i.e., correlation matrices) were reproduced from previous AFOQT investigations, all for samples of USAF pilot students. The first data set was originally analyzed in Duke and Ree (1996) with the purpose of showing that cognitive ability testing, as indexed by AFOQT and PCSM scores, plays a vital role in cost savings incurred as a result of selecting pilot candidates with a greater likelihood of training success. Descriptive statistics and correlations were the main procedures used in the study. The correlation matrix of the second data set was reproduced from Olea and Ree's (1994) study that assessed the incremental validity of specific abilities above and beyond the score of *g*. The *g* score was the first component extracted from a principal component analysis (PCA) of the 16 AFOQT subtest scores, while the specific ability represented by the remaining 15 components. PCA and regression analysis were the analytic techniques used in the study. Of the six criteria reported in the study, three were used in the cross-validation study of this thesis and the other three were used in the cross-occupation validation study. The third data set was obtained from Carretta and Ree (1997), which involved two samples of male and female student pilots. This study was an extension to Ree et al.'s (1995) study for the overall sample. Both the original and the extended studies had the goal of assessing a causal model relating general cognitive ability and prior job knowledge with job-knowledge acquisition and work-sample performance during training. SEM procedures were applied in the studies to achieve the targeted objectives. The predictive model assessed in the current thesis differs from that tested in the original study.

The cross-occupation validation study in the thesis involved three separate AFOQT samples: from pilots, navigators, and air battle managers. Correlation matrices for the pilot and navigator samples were reproduced from Olea and Ree (1994), whereas that for the air battle manager sample was reproduced from Carretta (2008). The primary goal of Olea and Ree's (1994) study was to demonstrate that general ability is the factor accounting for much of the variance in performance criteria of pilots and navigators, as compared to specific abilities. As mentioned above, a PCA was applied to extract 16 components for the general (first component) and specific (15 component) abilities. Carretta's (2008) study aimed at validating the 11 subtests of AFOQT Form S for predicting performance in air battle management jobs. Both studies involved correlation and regression analyses, in addition to PCA

Except for Carretta and Ree (1997), the rest of the studies did not conduct any SEM analyses for relating variables, and the results are only appropriate for interpretation at the observed score level. Even in Carretta and Ree's (1997) study, the only SEM-based study among the illustrated ones, the covered goals and research questions differed from those intended in the current thesis. In light of this, the four investigations considered in the present thesis are unique in their scope, goals, and research questions, as well as in their analytical procedures. They are conducted with a methodology that can be insightful for assessing the relations between cognitive abilities and flying performance. Beyond that, the data sets selected for this thesis are valuable because they were collected over many years and represent a population(s) with unique characteristics, from which sourcing information for a large sample is not always viable. Hence, it is beneficial to make use of such difficult-to-source data in new inquiries as long as they add value to the literature and are utilized in a new direction of research.

2.3.8 Modeling Job Knowledge Tests in AFOQT

The contemporary models of intelligence such as the CHC model include a broad factor for domain-specific knowledge (Gkn) to encompass constructs of specific and specialized knowledge. Domain-specific knowledge is defined as "the depth, breadth, and mastery of specialized declarative and procedural knowledge (knowledge not all members of a society are expected to have)" (Schneider & Mcgrew, 2018, p. 117). Such specialized knowledge is typically acquired through one's career, hobby, or interests. Job knowledge tests that are often added to selection test batteries are examples of domain-specific knowledge tests. Hunter (1986) documented a strong relationship between *g* and job knowledge and between job knowledge and performance. Job knowledge is seen as a mechanism through which cognitive ability influences job performance (Schmidt et al., 1986).The typical modeling of job knowledge in training involves causal relations with *g* in a manner that *g* affects the acquired job knowledge, which, in turn, affects job performance (Ree et al., 1995; Schmidt et al., 1986). Palumbo, Miller, Shalin, and Steele-Johnson (2005) showed that job knowledge tests outperformed cognitive ability tests directly as a predictor of task performance and indirectly as a mediator for the effects of cognitive ability on task performance.

The AFOQT factor analytic investigations frequently reveal a specific factor for job knowledge (or aircrew aptitude/interest or technical knowledge) represented by three subtests: AI, IC, and GS. The subtests in this factor are typically viewed and discussed differently than other ability-based subtests (e.g., Ree et al., 1995). In this thesis, I view job knowledge subtests (AI and IC tests) as ability tests measuring cognitive constructs that are relatively highly saturated with the broad factor of Acquired Knowledge. Such tests, nonetheless, are not expected to be as strongly correlated with other factors or as highly-loaded on *g* factor as those of Reading Comprehension or Arithmetic Reasoning, for example, which have a well-established pattern of high *g*-loading. Stated differently, job knowledge tests share common variance with typical cognitive ability tests but also have a considerable amount of unique variance.

2.4 The Great Debate on the Relative Importance of Cognitive Abilities

The relative importance of general cognitive ability versus specific abilities for job performance has been a subject of great debate (Kell & Lang, 2018). Despite the assertion that cognitive abilities are among the best predictors of job performance, the controversy as to which ability or set of abilities play a significant role in explaining the variance in the criterion space of job performance has never ended. In this context, dozens of studies have been and are being published to advocate the importance of particular ability (or abilities) over other abilities for job performance prediction. The works of Schmidt and Hunter (1998; 2004) are a strong line of research ascertaining that *g* is the most crucial ability for predicting performance in occupations, whereas specific abilities do not explain much variance beyond *g*. Hunter (1986, p. 341) took an extreme position when he claimed that “it is general cognitive ability and not specific cognitive aptitudes which predict job performance.” Schmidt (2002) argued that it is “not logically possible” (p. 187) to have a serious debate over the importance of general cognitive ability for job performance. In a similar vein, the “Not Much More Than *g*” series of Ree and his colleagues (Ree & Earles, 1991; 1996; Ree et al., 1994) is a reflection of the same standpoint that views *g* as the best construct for the prediction of job performance. One implication of such a hypothesis is that the focus in selection procedures should be directed, to no small extent, to applicants’ scores of general ability, *g* (or IQ), and to a much lesser extent toward their narrower ability scores.

Contrasting with this line of cognitive ability research, another direction that has been gaining attention in recent years emphasizes that specific abilities can also be significant components for predicting success in occupations, and their roles should not be ignored (e.g., Goertz et al., 2014; Murphy, 2017; Reeve et al., 2015; Schneider & Newman, 2015; Wee et al., 2014). The idea of having one single trait, *g*, capable of fully capturing the individual differences in job performance might be problematic for I/O psychology, particularly for selection and assessment purposes. Three challenges arise when relying solely on *g* score: violation of legal frameworks in some organizations (e.g., not complying with job analysis), limitations of information obtained from one single score, and the large majority-minority mean differences typically associated with *g* scores (Beier et al., 2019). It has been criticized that research examining the prediction of job performance often takes the value of *g* for granted, and other abilities are considered only for the sake of a little improvement. Stankov (2017) argued that the *overemphasized “g”* has hindered the study of broad and specific cognitive abilities and led to neglecting the first- and second-stratum factors in the CHC model. Similarly, Murphy (2017) noted that studies stressing *g* measures over measures of specific abilities fail to consider the second-stratum abilities that can sometimes be more predictive for job performance than global measures of general cognitive ability.

In contrast to the "Not Much More Than *g*" hypothesis, Kell and Lang (2017) showed that specific abilities in some workplaces could even be "More Important Than *g*." The supporters of this contention believe that many of the findings that have devalued the significance of specific abilities in workplaces were partially due to the limitations in those studies' analytical procedures. The majority relied primarily on regression-based analyses, which might not be the ideal analyses for making a firm conclusion about the relative importance of predictors. Although this family of statistical techniques is powerful in maximizing the prediction of a particular set of variables, they tend to provide an "unequal" opportunity for predictors to exhibit their potential power, especially when the multicollinearity among predictors is high (Tonidandel & LeBreton, 2011).

In hierarchical regression analyses, the most frequently used method in incremental validity studies, a score of *g* (often the first unrotated principal component or composite score from a test battery) is entered first in the model, whereas specific abilities are added second in the model (e.g., Ree et al., 1994). Criterion scores (e.g., flying performance) in such analysis are regressed first on scores of *g*, with scores of specific abilities (e.g., spatial ability, perceptual speed) entered in the second step of the regression. The shared variance in this statistical design is always attributed to the influence of *g* because the model prioritizes predictors entered first into the hierarchical regression, regardless of specific abilities' variance shared with the criterion. Even the overlapping shared variance between *g* and specific abilities is counted as resulting from *g*. The only variance that is credited to other predictors in the model is the percentage that does not overlap with *g*. Such an analytical strategy is likely to leave little remaining variance in criterion scores that can be accounted for by specific abilities (Lang et al., 2010).

For that reason, many researchers have called for adopting better analytical strategies when attempting to establish whether specific abilities have incremental validity above and beyond that provided by *g*. Murphy (2017) cautioned that the many publications overstressing the prediction role of *g* and understating the incremental contribution of specific abilities might have led to a premature decline in research on the roles of specific abilities in the workplace. Coyle (2014; 2018) postulated that some specific abilities could be found of equal or even higher importance than *g* in predicting outcomes when the relations are tested via the non-*g* residuals of test scores. He regarded this method as the most promising approach in the study of human intelligence. Contrary to the primacy of *g* hypothesis, he was able to uncover significant incremental validity for many specific abilities on the SAT, ACT, and the Preliminary SAT tests above *g* validity for the prediction of diverse criteria, often with substantial effect sizes ($\beta s \approx 0.30$) (for review, see Cole, 2018).

The *bandwidth-fidelity dilemma* (e.g., Ones & Viswesvaran, 1996; Salgado, 2017) is another concern frequently raised in investigations of predictor-criterion relationships and is believed to be one factor biasing against showing appreciable effects of specific abilities. The center point here is the necessity to make a reasonable alignment between predictors and criteria such that general predictor is matched with the general criterion, and specific predictors are matched with specific criteria. Schneider and Newman (2015) labeled this strategy as the *ability–performance compatibility principle* to show that general abilities predict general job performance, while specific abilities predict specific job performance. Drasgow (2012) argued that expanding the criterion space to include criteria other than training performance and overall job performance (e.g., contextual job performance, counterproductive work behaviors (CWBs), and attrition) might be another way to better understand the individual differences that predict behavior in the workplace. Hence, a more thoughtful plan in the design and implementation of validation research can have an impact on the conclusions that can be derived about the true relations between variables involved in the study.

The recent study of Wee (2018) can be a good illustrative example as she addressed both concerns mentioned above; that related to predictor-criterion bandwidths and that associated with the analytical procedure. She planned her study to include different breadths of cognitive predictors (general ability and three specific abilities [spatial reasoning ability, verbal reasoning, numerical reasoning]) and varied breadth of performance criteria (general performance and four specific performances [Math, German, English, Sports]). The predictive relations were tested using two statistical techniques: SEM and relative weights analysis. Results revealed that the relative importance of general and specific abilities varied with the analytic procedure used. Based on the SEM approach, none of the specific abilities had the power to predict either general or specific performance. General ability seemed to be a significant predictor for general performance, but not for specific performance. In contrast, based on relative importance analysis, results showed that verbal reasoning predicted general academic performance more strongly than general ability or any other specific abilities, and also, it predicted each of the specific subject grades better than any other ability. Although the study failed to support the value of predictor-criterion alignment, it did provide evidence supporting the utility of specific abilities, in addition to *g*, as useful predictors of performance.

The bright side of this long-lived scientific debate, however, is that it has stimulated dynamic research in both directions, which is certainly advantageous for the advancement of related sciences. Some journals have devoted special issues to debating the relative value of cognitive abilities for performance outcomes. As an example, a special issue of *Human Performance* included 12 articles discussing the role of general mental ability in I/O psychology

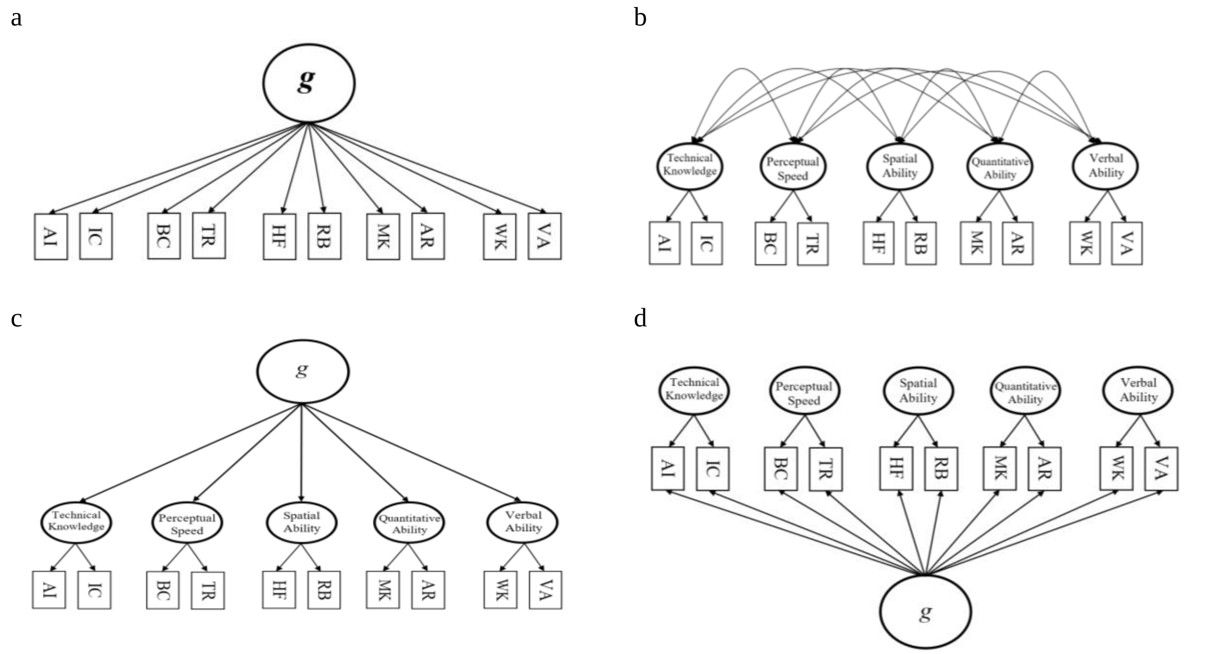
(Viswesvaran & Ones, 2002). Equally, the recent special issue of the *Journal of Intelligence* focused on this *great debate* in seven articles (Kell & Lang, 2018) in an attempt to motivate a reconsideration of specific abilities in the workplace. Some of these articles have offered useful analytical strategies that can be used as an alternative to the traditional statistical analysis to disclose determinants of job performance (e.g., Coyle, 2018; Eid et al., 2018; Ziegler, & Peikert, 2018). Of interest, this debate on the relative role of general versus specific abilities has transferred from educational and workplace settings to other life domains. Some forms of this debate can now be found in studies of wages (Ganzach & Patel, 2018), players of the National Football League (Lyons et al., 2009), happiness (Blasco-Belled et al., 2019), triangular love (Van Buskirk, 2018), humor production ability (Christensen et al., 2018), music training (Silvia et al., 2016), and piano skill acquisition (Burgoyne et al., 2019).

2.5 Modeling Cognitive Abilities Structure

Intelligence theories are commonly described via statistical models specifying their posited structures. Some models are more indicative of particular theories than others. For example, the single-factor model is a good representation of Spearman's theory (Spearman, 1904), while the higher-order factor model with its three levels of abstraction is more indicative of Carrol's three stratum theory (1993). Attempting different modeling techniques when assessing the factor structure of cognitive data can be useful, and, in fact, is a recommended strategy (Jöreskog & & Sörbom, 1993). Although no agreed-upon model has yet been reached for the best conceptualization of cognitive ability structure, theoretical considerations usually guide researchers' choices on how the data are best modeled. As observed by Molenaar (2016), for example, Horn's (1968) arguments were in favor of the correlated factor model, Jensen's (1998) and Johnson and Bouchard's (2005) arguments were in support of the higher-order factor model, and Gignac's (2008) and Beaujean's (2015) arguments were in favor of the bifactor model. In a comparison between three statistical models (i.e., correlated-factor, higher-order, and bifactor models), Morgan et al. (2015) recognized a wide overlap of fit values across the compared models irrespective of true structure from which they were obtained and thus, they suggested to judge the models on substantive and conceptual grounds.

Relatedly, while there is relative agreement about the psychometric existence, stability, and significance of *g*, there is as yet no consensus about the factor structure beyond *g* (Valerius & Sparfeldt, 2014). Given the focal role of *g* in cognitive research, it is critical to understand how *g* is conceptualized and extracted. Many procedures have been suggested to derive *g* scores from cognitive test batteries (e.g., Jensen, 1998; Reeve & Blacksmith, 2009). The contention from some studies is that the *g* estimate is likely stable across different methods of factor extraction (Jensen & Weng, 1994; Ree & Earles, 1991; Thorndike, 1987). More recently, Floyd et al. (2009) examined the influence of three design features on identification of the general factor (i.e., the factor extraction method, the composition of test batteries, and the number of tests in the battery. After combining many different configurations of test batteries and factor extraction, generalizability theory analyses revealed a dependability coefficient for general-factor loadings of .99 and that the characteristics of the tests had the largest percentage of variance contribution.

In this thesis, I used SEM procedures to assess the relative contribution of *g* and specific ability factors in *predicting* pilot training performance. Conclusions about the role of cognitive abilities in flight performance have mostly been established using statistical techniques more suitable for observed variables (e.g., correlation and regression analysis). Hence, predictive validity studies can benefit from SEM procedures because they capture the underlying factors contributing to test performance (Muthén, 2015). In selection research, the interest generally goes beyond the observed scores of test batteries (often associated with measurement error) to the relationship between the "true" ability constructs and the true performance outcome. Understanding the relationships between variables at a more abstract level (i.e., true-score variables) is an often desirable goal. Researchers make effort to move from the observed world of variables to the theoretical world of latent variables, which enables creating a more definite conclusion about what ability(s) (e.g., verbal, quantitative, spatial) should be considered in the selection strategy. Besides, due to the likely influence of analytical procedures on the results, it would also be of interest to attempt SEM for assessing how conclusions may differ as a function of the analytical methods used. Below is an explication of four CFA models (single-factor, correlated factor, higher-order factor, and bifactor) commonly used to describe cognitive data obtained from test batteries. Representation of these four models is provided in Figures 2a, 2b, 2c, and 2d, respectively, which mirror the factors and subtests used in Study 4 of this thesis. The extension of the CFA models to SEM models by specifying paths from cognitive factors (predictors) to performance measures (criteria) will also be pointed out. The next section will include more about SEM as a methodology featuring powerful capability in modeling predictive relations.



Note. VA = Verbal Analogies; WK = Word Knowledge; AR = Arithmetic Reasoning; MK = Math Knowledge; RB = Rotated Blocks; HF = Hidden Figures; TR = Table Reading; BC = Block Counting; IR = Instrument Comprehension; AI = Aviation Information.

Figure 2. CFA models. (a) Single-factor mode; (b) Correlated-factor model; (c) Higher-order factor model; (d) Bifactor model

2.5.1 Single-factor Model

The single-factor model is one of the simplest and oldest conceptualizations of human mental structure, which mirrors Spearman's theory of the two-factor model (Spearman, 1904). The underlying assumption in this model is that cognitive data (i.e., ability test scores) are influenced by a single latent ability factor shaping the performance in all cognitive tasks and accounting for the shared variance between observed indicators (Kline, 2015; Spearman, 1904). This indicates that individual differences in ability tests are mostly due to the effect of *g* rather than any specific abilities. Theoretically, the single-factor model suggests one general construct (or *g*), causing the indicators to inter-correlate with each other. For data obtained from test batteries, this model rarely yields acceptable goodness-of-fit statistics (Schneider & Newman, 2015), which indicates that the variance in cognitive tests cannot be explained by one single ability. If this model is tenable, however, it can be extended to a predictive model by specifying a link from the latent factor of the general ability to a performance measure. The goal in such a predictive model is to estimate the effect of *g* on performance criteria to assess its predictive value. Only one regression effect is produced through this model.

2.5.2 Correlated-factor Model

The correlated factors model (or correlated-trait model) corresponds to intelligence models that deemphasize the existence of e general ability influencing performance in all cognitive tasks. Thurstone's model of primary mental abilities (Thurstone, 1938) and the extended Catell-Horn model of Fluid–Crystallized (*Gf*-*Gc*) Intelligence (Horn, 1991) reflects some of the underlying assumptions of this model. This model is more complex than the single-factor model. Also, it differs from the single-factor model in the sense that it assumes the variability in the performance of cognitive tests is best accounted for by a group of correlated ability factors, rather than one single ability. It implies that the variance common to all indicators is due to a shared set of latent factors instead of one general factor.

Model comparison between a single-factor model and a correlated-factor model is often recommended (Brown, 2015; Byrne, 2013). The extended predictive model of the correlated-factor model is specified by linking the ability factors with performance measures to estimate their predictive power. The effects are estimated by regressing performance scores on the scores of predictor variables. The performance in this model is seen as a function of the correlated abilities that are the determining factor of performance. Similar to multiple regression analysis, the estimated regression coefficients represent the marginal effect of each ability on performance, with the other ability is held constant. The number of effects (regression coefficients) produced is equal to the number of predictor factors. Nevertheless, Gignaca and Kretzschmarb (2017) cautioned against taking the results from this model, on their own, as evidence for separability, or uniqueness, of each of the hypothesized specific factors. Instead, they recommended to further analyze the data using a higher-order model.

2.5.3 Higher-order Factor Model

This model assumes that the representation of cognitive data takes a hierarchical structure, with *g* at the top and the ability group factors below, and specific abilities, as indicated by observed test scores, at the lowest level. The fundamental assumption in these models is that both *g* factor and ability group factors contribute to the performance in cognitive tasks. However, the effect of *g* on ability subtests is indirect, whereas the effects of the group factors are direct. That is, such a representation implies that the lower-order factors of specific abilities fully mediate the influence of *g* on task performance; thus, *g* influences the subtests indicators only indirectly and operates only through the lower-order factors. The appropriateness of hierarchical models to describe human cognitive ability structure as opposed to the one-factor model has been well-established (e.g., Gustafsson, 2001). Reeve and Blacksmith (2009) found that almost 57.6% of the articles applying CFA methods have used a higher-order modeling strategy.

Recent use of this model for the assessment of predictive relations show three specification variants (e.g., Benson et al., 2016; Berkowitz & Stern, 2018; Coyle, 2018; Wee, 2018). In the first variant, the *g* factor is the only factor linked with performance criteria, and hence, the outcome is only one regression coefficient representing *g* effect. The performance score is assumed to be influenced primarily by *g*, despite the existence of specific abilities' effects. The second variant involves adding paths from specific abilities to performance measures parallel to that from *g* to performance measures. The influence of specific abilities on performance thus can be estimated, and the effect of each ability on performance, controlling for the other abilities in the model, can be revealed. Another way to account for the impact of specific abilities is by allowing their residuals to be correlated with performance measures, a technique that results in partialling out the effect of *g* and provides a somewhat purer estimate of specific abilities effects on performance. The third version is similar to the first version, but here the residuals of specific factors (non-*g* residuals) are allowed to be correlated with the criterion. This method has recently seen increased use and has assisted in establishing the predictive validity of several specific constructs beyond the validity obtained by *g* factor (e.g., Coyle, 2014; 2018).

2.5.4 Bifactor Model

The bifactor or nested-factor model was introduced many decades ago by Holzinger and Swineford (1937). However, its use as a predictive model for the associations between predictors (e.g., cognitive abilities) and outcomes criteria (e.g., job performance) has only recently been revived. The unavailability of this model in commonly-used statistical software may have delayed its applications in predictive relations. Presently, bifactor models are used widely for defining the relationships between cognitive constructs. The readily built-in *orthogonalization* feature in this model makes it appropriate for investigations that seek a complete distinction between the effects of general and specific factors, such as those in this thesis. The effect of general and specific factors on criterion variables thus can be examined through latent multiple regression models underlying the SEM algorithm.

The bifactor model is less constrained than the higher-order model and also has distinct statistical specifications. In the bifactor model, *g* is modeled similarly to specific abilities as a lower-order factor, but different in that it has paths to all (or the majority) of the indicators, instead of only a specific group of indicators. Both *g* factor and specific ability factors work independently and influence test performance separately from each other (e.g., Brunner et al., 2012). Because the higher-order model is a bifactor model imposed with more constraints, it is mathematically nested within the bifactor model. Studies comparing the two models found that bifactor models produce a better fit than the higher-order models (e.g., Cucina & Byle, 2017). Morgan et al. (2015) also indicated that the higher-order model tends to produce a weaker fit than the bifactor model, even when samples are obtained from a true higher-order structure data. Simulation results show that when there is unmodelled complexity (e.g., correlated residuals and cross-loadings), the CFA results tend to favor the bifactor model over the higher-order model (Murray & Johnson, 2013). The examination of predictive relations in the present thesis relied primarily on a bifactor modeling approach as a better alternative than higher-order models for assessing the interplay of general ability and specific abilities in predicting job performance.

2.6 Structural Equation Modeling (SEM)

SEM, as defined by Ullman and Bentler (2012; p. 661) is "a collection of statistical techniques that allow a set of relationships between one or more independent variables (IVs), either continuous or discrete, and one or more dependent variables (DVs), either continuous or discrete, to be examined." CFA and path analysis are special cases of SEM. The typical SEM model contains two parts, a *measurement* model representing the relationship between the latent variables and their measured indicators and a *structural* model describing the relationship between the latent (or observed) variables of interest (Ullman & Bentler, 2012). SEM procedures can be used for testing measurement, functional, predictive, or causal hypotheses (Bagozzi and Yi, 2012).

SEM can handle models of complex structures and address the research problem of multivariate causal hypotheses (Marcoulides & Yuan, 2017). As compared to a simple single or multiple linear regression, SEM is known for its flexibility and capability to analyze a system of regression equations (Schumacker & Lomax, 2004). It operates several equations and test all parameters *simultaneously*, as compared to the ordinary regression analysis that considers each equation separately. Via SEM, decomposition of observed variables (or their variances and covariances) into true scores and errors is achievable (Nachtigall et al., 2003).

SEM allows comparison between the theorized model and empirical data, through several fit-statistics showing the degree of conformity between model and data. It also enables testing different models, with the possibility of making a comparison between any nested models (Kline, 2015). An SEM model is judged to be supported by the data if the resulting fit-statistics are found to meet the acceptable level of fit. The adequacy of fit statistics suggests that the empirical examination (the collected data) supports the theorized relationships between latent and observed variables (measurement models) and that between the latent variables (structural model). However, Nachtigall et al. (2003) cautioned that a well-fitting SEM model does not necessarily imply causal dependencies, although, under certain circumstances, SEM may embody causal relationships. Thus, it is important to understand that testing the fit of an SEM is not a test of causality and that SEM by themselves do not demonstrate causality (Bagozzi & Yi, 2012).

Due to SEM strength in modeling relationships between latent variables, it can be appropriate procedure for the different goals intended in this thesis. The two-step approach suggested by Anderson and Gerbing (1988) is implemented, where the measurement models are established first, and the latent paths models are being introduced accordingly. Regarding the measurement models of SEM used in the four studies of this thesis, a general strategy for specifying latent construct is to include at least two indicators per factor to ensure an acceptable level of construct validity (i.e., the extent to which indicators of a construct (or factor) measure what they are purported to measure). Hayduk and Littvay (2012) recommend using fewer best indicators over the multiple indicators and argued that two indicators could be sufficient, although three indicators may be helpful.

2.7 Summary

Flying is a demanding job that involves multiple skills and aptitudes. It is also associated with a high cost in training and a high risk to human life. Thus, selecting pilot applicants with high potential of success is an appealing goal that every flying organization seeks. Multiple phases of selection is perhaps the most preferred procedure for selecting pilot applicants, which entails diverse selection methods and tools including cognitive and psychomotor test batteries, structured interview, work sample, assessment center, as well as personality and biodata assessment. Cognitive testing, in particular, has seen popularity in selection settings due to its relatively low cost and the feasibility of administration for a large number of applicants. The AFOQT is one of these test batteries that has been in use by the USAF since the emergence of the psychometric testing movement after World War-II. The long history of this instrument and the large number of validation studies reported for its observed scores of pilot performance make it appropriate for further assessment at a more abstract level. This thesis reports four studies conducted with several AFOQT datasets targeting a broad goal of assessing the role of cognitive abilities in predicting future success of pilots in their training program. The SEM approach was found to be appropriate for the different objectives posited for the four investigations. The outcomes of this thesis can be of significance in the effort to thoroughly understand the human factors associated with flying occupations.

CHAPTER 3: RESEARCH METHOD

This chapter outlines the research methodology applied in this thesis. The preliminary study and the other three predictive studies are conducted with the main goal of identifying the role of particular cognitive abilities in the prediction of pilot performance. The use of SEM procedures and multiple data sets can support drawing, and properly delimiting, generalizable conclusions. Each study contributes uniquely to the objectives of this thesis.

3.1 Preliminary Study

Before carrying out the three predictive studies, I conducted a preliminary study using Cheung and Chan's (2005) approach to meta-analysis, meta-analytic structural equation modeling (MASEM). The use of MASEM methodology is advantageous in that it combines the strength of two powerful modeling techniques: meta-analysis and SEM. This approach offers a two-stage SEM (TSSEM) procedure; the first stage is designed to construct a pooled correlation matrix from the collected data, while the second stage is designed for testing a preidentified SEM model. This study only used the first stage of analysis to synthesize the intercorrelation matrices between subtest scores reported by prior investigations of the AFOQT. The resulting pooled correlation matrix was then used as input to conduct an exploratory factor analyses. The goal of this study was to examine the factor structure of the AFOQT by means of EFA procedures. More specifically, after obtaining the meta-analytic correlation matrix, I examined one- to six-factor EFA models. The key interest here was to investigate the ability constructs that may emerge from each solution and to note the relations between the subtests and constructs. This should give a solid basis for the predictive models used in the subsequent studies. Further, a second-order EFA with *Schmid-Leiman orthogonalization* procedure (Schmid & Leiman, 1957) was also applied in the study. This analysis provides further understanding of instruments presumed to measure higher-order factors or correlated traits (Dombrowski, 2015). However, it was solely used for the five-factor model, the structure suggested for AFOQT in previous studies. The combination of the two exploratory approaches can be useful and enhance the level of confidence in the modeling process in the studies. Because the main purpose of using the AFOQT in this thesis is to source data suitable for capturing broad ability factors, this preliminary study will be useful in providing evidence supporting their construct validities.

3.1.1 Search Strategy and Inclusion Criteria

The primary terms used for obtaining studies for the meta-analytic study were "*Air Force Officer Qualification Test*" and "*AFOQT*." Three main databases and search engines were used in this step: *The Defense Technical Information Center*, *Google Scholar*, and the general search engine of *Google*. The extracted studies underwent initial screening using the study's abstract and skimming through the paper. The main criteria pre-conditioned for studies' inclusion were that the primary study had to include at least one correlation coefficient between any pairs of AFOQT subtests, as well as reporting sample size. Additionally, a clear description had to be reported by the primary studies about participants, data collection, ability and performance measures, and analytic plan. The initial plan was to include in the intended meta-analysis all correlation matrices reported for AFOQT, regardless of the subpopulation of officer applicants to whom it was administered. However, comparing cognitive structures through initial EFA revealed differences between pilot samples and non-pilot samples (e.g., norm group). For example, the strength of intercorrelations and the estimate of psychometric *g* in pilots' data was found to be lower than that of normative data. Hence, it was decided to include in the meta-analysis only those studies that reported correlation matrices for pilots, which were relatively few (k = 11; n = 131 – 7563; N =

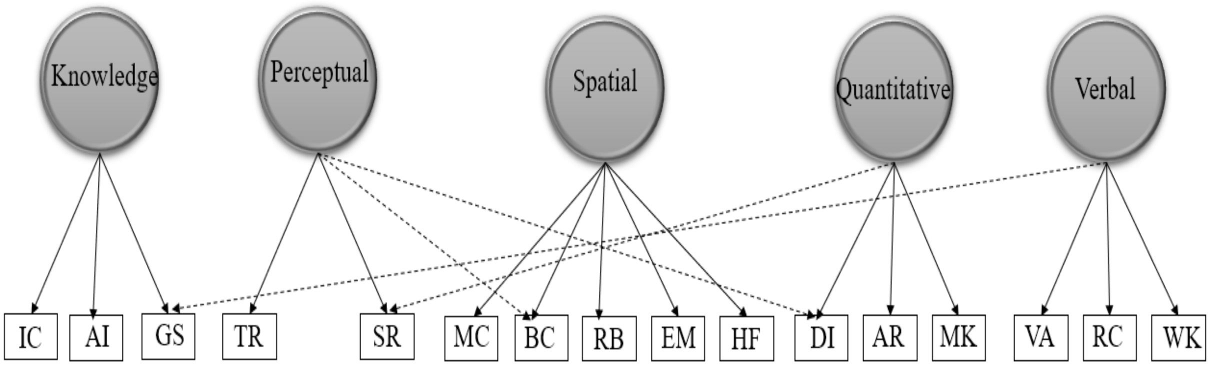
21102). Of the 11 correlation matrices deemed appropriate for the meta-analysis, five contained all 16 AFOQT subtests, one contained nine subtests (Carretta & Ree, 1997; VA, RC, WK, AR, DI, MK, SR, IC, and AI), and five contained a different set of nine subtests (Hunter & Thompson, 1978; VA, WK, MC, EM, SR, BC, TR, RB, and HF). Afterward, the approved 11 matriciabes for inclusion were reproduced and prearranged for the subsequent meta-analysis.

3.1.2 Meta-analysis Procedure

The available 11 correlation matrices were pooled together in a meta-analytic investigation. The Cheung and Chan's (2005) TSSEM approach was used as the method for synthesizing matrices from the primary studies. In TSSEM, a multiple-group CFA model is used to assess the homogeneity of the correlation matrices and to synthesize the matrices (Cheung & Chan, 2009), where groups in this context are primary studies. A random-effects model based on maximum likelihood (ML) estimation was applied to formulate a multivariate meta-analytic on the collected correlation matrices (e.g., Cheung & Cheung, 2016). Due to the missingness of some correlation coefficients in the collected data, an ML-based estimation method can be efficient in providing unbiased parameter estimates (Enders, 2010). Similarly, due to the heterogeneity expected between studies, a random-effects model can be more suitable than a fixed-effects model. Potential sources of heterogeneity in this study were the variation in AFOQT forms administered and the wide range of years of studies. The test of heterogeneity assesses whether the correlation matrices obtained from different primary studies can be assumed to be derived from the same population. The variability across study results beyond random sampling error is referred to as between-study variance (Veroniki et al., 2016). The quantification of between-study variance has crucial role in assessing the results interpretation of meta-analysis. To quantify the effect of heterogeneity, two statistics were used: Cochran's Q and I^2 (e.g., Higgins & Thompson, 2002). One Q estimate is produced for the overall degree of inconsistency in the studies' results, and a unique I^2 statistic is given for pairwise correlation to indicate the percentage of total variation across studies that is due to heterogeneity rather than chance. Thus, 120 total I^2 estimates were produced. A non-significant Q estimate and small percentage of statistically significant I^2 estimates (< 25%) may be taken as evidence of homogeneity.

3.1.3 EFA of the Meta-analytic Correlation Matrix

The resulting pooled correlation matrix was used as an input for factor-analytic examinations. To thoroughly understand the broad ability factors underlying AFOQT subtests, EFA was utilized in the assessment of one- to six-factor AFOQT structures. For a better comprehension of AFOQT structure at a higher-order level, the five-factor model, the suggested representation of AFOQT data, was followed by a second-order EFA with *Schmid-Leiman orthogonalization* (Schmid & Leiman, 1957) procedure. All EFAs conducted were based on *Principal Axis Factoring* (Fabrigar et al., 1999) and *Promax* rotation (Gorsuch, 1983). A path diagram for the theorized five-factor structure of the 16-subtest AFOQT (Carretta & Ree, 1996; Skinner & Ree, 1987) is presented in Figure 3.



Note. Solid lines represent paths from a theorized ability factor to its primary indicators; dashed lines represent paths from a theorized ability factor to its secondary indicators; VA = Verbal Analogies; AR = Arithmetic Reasoning; RC = Reading Comprehension; DI = Data Interpretation; WK = Word Knowledge; MK = Math Knowledge; MC = Mechanical Comprehension; EM = Electrical Maze; SR = Scale Reading; IC = Instrument Comprehension; BC = Block Counting; TR = Table Reading; AI = Aviation Information; RB = Rotated Blocks; GS = General Science; HF = Hidden Figures

Figure 3. Factor structures of the 16-subtest Air Force Officer Qualifying Test (AFOQT)

For a better judgment of the number of factors to be retained, multiple criteria were used including the Kaiser's (1960) mineigen greater than 1 criterion, Cattell's (1966) scree test, Horn's (1965) parallel analysis (HPA), Velicer's (1976) minimum average partial method (MAP), Bayesian information criterion (BIC; Schwarz, 1978), and sample size adjusted BIC (SSBIC; Sclove, 1987; see also, Morgan, 2015). MAP test involves two-step analysis; partialing out the correlation matrix via principal components analysis and computing the squared partial correlations. The point where the minimum average of the squared partial correlations is reached indicates the number of factors to retain (Garrido et al., 2011). For BIC and SSBIC, the model with the lower value, and a 10-point difference is considered superior (Kass & Raftery, 1995). The recently developed exploratory graph analysis (EGA; Golino & Epskamp, 2017) was also attempted as an additional criterion. EGA detects the dimensions of psychological constructs from the network psychometrics perspective. Psychometric network analysis conceptualizes psychological attributes as interconnected networks, with two main elements forming its structure; nodes and edges. Nodes represent the indicators (the observed scores), and the edges represent the connections between the nodes (pairs of indicators). The edges (or links) connecting the nodes are typically partial correlation coefficients between the variables (e.g., subtests) that show the strength of their association (Epskamp & Fried, 2018).

Utilizing this number of criteria (i.e., seven) should be informative for suggesting the most plausible AFOQT model. The factor solution of each model was assessed for interpretability and theoretical conceivability (Fabrigar et al., 1999), as well as three model fit indices: Tucker Lewis Index (TLI; Tucker & Lewis, 1973), the Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1992), and the BIC. To be considered a plausible solution, each factor had to be marked by two or more salient loadings and, preferably, no salient cross-loadings (Gorsuch, 1983). Salient factor pattern coefficients were defined as those loadings equal to or greater than .30 (Child, 2006). Following Hu and Bentler's (1999) recommendation, good (or acceptable) model fit requires values of TLI > 0.95 (> 0.90) and RMSEA < 0.06 (< 0.08). A model with lower BIC value is more preferred than a model with higher BIC value (Kass & Raftery, 1995).

3.2 Data Sets

The studies conducted in the present thesis have utilized data from different sources, all of which include correlation matrices for AFOQT scores and different pilot performance measures (i.e., Carretta & Ree, 1995; 1997; Duke & Ree, 1996; Olea & Ree, 1994). Additionally, two non-pilot samples representing navigators and air battle managers were also obtained (Carretta, 2008; Olea & Ree, 1994) so as to be used in the cross-occupation validation study. These studies were selected from the pool of studies collected for the meta-analytic preliminary study (11 datasets) because they were the most recent studies that included a variety of performance measures (3 to 17) and large sample sizes (N > 950). The air battle managers sample was the only sample that had a single performance measure and was slightly smaller in size (N = 680), although still large enough. The versions of AFOQT administered to all samples were the 16-subtest AFOQT (Form O, P, or Q), with the exception of the air battle manager sample that represented Form S. Also, one of the three datasets utilized in the cross-validation study reported correlations for only 9 of the 16 subtests, which limited the number of factors possible for extraction. Following is a brief description about the six studies from which the datasets were sourced, grouped by the study in which they are going to be utilized in this thesis.

3.2.1 Primary Validation Study

The primary predictive study in this thesis used a data set that was reported and assessed in two prior published studies: Carretta and Ree (1995) and Johnson et al. (2017). The sample included 7,563 undergraduate pilot training (UPT) pilot students. The 16 subtest scores of the AFOQT were all reported in the study. The performance criteria used in this study were five rating scores collected during three phases of training, four of which were actual flying performance during primary and advanced phases of training and one academic performance.

3.2.2 Cross-validation Study

3.2.2.1 Sample 1

The correlation matrix of the first data set used in the cross-validation study was directly reproduced from Duke and Ree's (1996) study. The sample included 1,082 UPT pilot students. Data for all 16 AFOQT subtests were available. Three performance measures were reported in the study indicating performances in primary and advanced phases of training, as well as one overall score.

3.2.2.2 Sample 2

The second data set used in this investigation was sourced from Olea and Ree's (1994) study, which reported two data sets for pilots and navigators. The pilot sample consisted of 1,867 undergraduate pilot training students. The navigator sample included 957 undergraduate navigator training students. For each sample, data were available for 16 AFOQT scores and 6 performance scores. Three of the pilots' performance measures were utilized in this study, and the other three were reserved for the cross-occupation validation study.

3.2.2.3 Sample 3

The correlation matrix of the third data set was reproduced from Carretta and Ree's (1997) technical report. Two data sets were reported in the study, one for male pilots (N = 3,369) and the other for female pilots (N = 59). The initial plan was to apply multigroup SEM to assess whether the suggested predictive model can be applied for both genders. However, due to the severely unbalanced samples (e.g., Yoon & Lai, 2018) with almost 1:57 ratio, I used in the study only male pilot data, without attempting any multigroup SEM. The correlation matrix contained data for 9 AFOQT subtests and 17 performance measures. The subtests assessed in the study included VA, RC, WK (verbal ability), AR, DI, MK (quantitative ability), IC and AI (acquired aviation knowledge), as well as SR (perceptual speed). Hence, constructs of spatial ability and perceptual speed were not represented adequately in the study.

3.2.3 Cross-occupation Validation Study

3.2.3.1 Pilots and Navigators

As mentioned above, Olea and Ree's (1994) study reported two data sets for pilots and navigators, both of which were used in the cross-occupation validation study. The pilot sample consisted of 1,867 UPT students, and navigator sample included 957 undergraduate navigator training students. Data were available for all 16 AFOQT subtest scores and six performance measures for each sample, three measures were used in this study.

3.2.3.2 Air Battle Managers

The data of air battle managers was obtained from Carretta’s (2008) technical report. The sample consisted of 680 undergraduate air battle manager training students. The AFOQT used in this study was Form S that consisted of 11 subtests. The performance criterion in this sample was one overall index given to the students representing an average final score on several written tests taken during the training course.

3.3 Subjects

Pilots’ samples in the studies represented pilot trainees attending undergraduate pilot training (UPT) in the USAF. This program typically consists of three main phases: ground school, primary phase of training, and advanced phase of training. The AFOQT was the primary selection tool used in qualifying the subjects for officer training programs. Subjects in the samples had completed at least a four-year baccalaureate degree before training. The time elapsed between cognitive testing for officer selection and criterion data collection was between one to five years. Samples were dominated by whites (> 96%) and males (> 98%), with age typically in the range of 22 to 27 years upon completion of training. In addition to qualification on the basis of AFOQT scores, the selected applicants had to meet other selection standards such as academic achievement, medical, moral and physical fitness, personal recommendations, and prior flying experience. The cross-occupation study also included a navigators’ sample and an air battle managers’ sample, with similar characteristics to pilot samples, although they differed in the AFOQT composites with which were used to qualify them for recruitment (e.g., pilot composite, navigator composite).

3.4 Measures

3.4.1 Cognitive Abilities

Cognitive abilities in the three predictive studies (Studies 2, 3, and 4) were indicated by AFOQT subtest scores. Chapter 2 included an adequate overview of the AFOQT and its factor structure (see Table 2). Because a model with five specific ability factors is the most documented structure for AFOQT data, the extended predictive SEM studies will use the same model for cognitive abilities, as long as this structure is supported by the EFA in the preliminary meta-analytic study. However, the data available for the AFOQT subtests in two particular studies (Carretta, 2008; Carretta & Ree, 1997) may limit the extraction of all five ability factors. As a reminder, Table 3 presents the configuration of the AFOQT subtests, grouped by the theorized five-factor model (see also Carretta & Ree, 1996). Table 4 has information about the performance measures, but also includes the number of AFOQT subtests administered in each primary study.

Table 3. *The Five-factor Model Underlying the 16-test AFOQT and Subtests' Composition*

Latent Ability	**Subtests' Composition**
Verbal ability	Verbal Analogies (VA), Reading Comprehension (RC), Word Knowledge (WK), *General Science (GS)*
Quantitative ability	Arithmetic Reasoning (AR), Data Interpretation (DI), Math Knowledge (MK), *Scale Reading (SR)*
Spatial ability	Mechanical Comprehension (MC), Electrical Maze (EM), Block Counting (BC), Rotated Blocks (RB), Hidden Figures (HF)
Perceptual speed	Table Reading (TR), Scale Reading (SR), *Block Counting (BC), Data Interpretation (DI)*
Aviation acquired knowledge	Instrument Comprehension (IC), Aviation Information (AI), General Science (GS)

Note. The *italicized* subtests are those proposed to have secondary loadings on another ability factors, in addition to their primary loadings (unitalicized)

3.4.2 Pilot Performance

The measures used to indicate pilot performance varied across studies. Also, the number of performance measures used to estimate criterion-related validities in these studies differed notably. Table 4 details the type and the number of performance measures utilized in the studies collected for the purpose of this thesis. As shown in the table, some studies used only one performance criterion (Carretta, 2008), while others had as many as 17 measures (Carretta & Ree, 1997). This number and variety of performance measures will be useful to draw a more reliable conclusion about the role of abilities for various types of pilot (and the other two jobs') performances.

Table 4. *AFOQT Subtest and Performance Measures Reported in the Studies from Which the Correlation Matrices were Reproduced*

		Reported Subtests	Reported Performance	Performance Measures
Pilot	Carretta and Ree (1995)	16	5	(1) Phase 1 academic grade (2) primary phase daily flight ratings (3) primary phase check flight ratings (4) advanced phase daily flight ratings (5) advanced phase check flight ratings
	Duke and Ree (1996)	16	3	(1) The rank of students in the UPT class, (2) Extra flying hours in the basic phase, (3) Extra flying hours in the advance phase
	Olea and Ree (1994)	16	6	(1) Pass/Fail undergraduate pilot training (2) primary phase check flight ratings, (3) advanced phase check flight ratings, (4) overall composite for primary phase, (5) overall composite for advanced phase, (6) overall composite for the entire measures in the study
	Carretta and Ree (1997)	9	17	(1) three check flight ratings during the primary phase of training, (2) three check flight ratings during the advanced phase of training, (3) 11 end-of-course academic grades representing early (4 tests' scores), middle (4 tests' scores), and late (3 tests' scores) training
Navigators	Olea and Ree (1994)	16	5	(1) Pass/Fail undergraduate navigator training, (2) airmanship, (3) basic procedures, (4) day celestial check flight, (5) night celestial check flight, (6) overall composite for all five measures
Air Battle Management	Carretta (2008)	11	1	(1) Average final score of written exams for topics taught during training.

3.5. Analytic Procedure

A similar analytic procedure was planned for all three predictive studies. A total of seven analysis series were carried out to investigate the predictive relations between abilities and performance, one for the primary validation, three for the cross-validation, and three for the cross-occupation validation study. The result of the meta-analytic factor analysis will inform the modeling process of the predictive studies by supporting (or rejecting) the viability of the targeted five-factor model of AFOQT cognitive structure and also, by identifying the subtests that load highly on each factor. Following is the four-step analytic plan applied in each of the seven analyses.

3.5.1 Modeling Performance

A decision whether to model performance measures as latent or observed was necessary before applying SEM models. This was informed by the number of performance measures available in the data, their orientation (actual flying performance, academic performance), and the phases in which they were collected (primary, advanced). There was a preference for modeling performance as latent if there were a sufficient number of indicators. For example, the primary validation study reported five performance measures; two of which were indicative of flying performance during the primary phase of training and another two were indicative of the flying performance during the advanced phase of training, while the fifth measure were academic performance. Hence, the flying performance on both the primary and advanced phases of training were modeled as two latent variables each indicated by two observed measures. The academic performance was kept as observed. When the number of performance measures was large enough, EFA was conducted to verify their best representation, either one single construct or multiple constructs (only one case). For instance, the academic dimension of performance in Sample 3 of the cross-validation study had total of 11 measures. The EFA examination of the correlations between these 11 measures indicated that they are best represented by a single-factor structure; thus, they were configured as a latent performance factor for the purpose of SEM study.

3.5.2 Modeling Cognitive Abilities

Testing the construct validity of the measurement model of abilities was essential. A CFA bifactor model, as well as a correlated-factor model, were tested in this step. It was necessary to evaluate whether the cognitive abilities formed a coherent bifactor and correlated-factor structure, since they both will be used for assessing the predictive validity of cognitive abilities.

3.5.3 Correlation-based Validity

In this phase of analysis, the ability factors were associated with performance measures (latent or observed) in a correlated-factor models to assess how they correlate with each other. This initial assessment can be important in view of the fact that the dominant statistical analysis in criterion-related validity studies remains simple correlation. Moreover, this analysis will provide a frame of reference for comparing the validity of specific abilities before and after accounting for the domain-general factor.

3.5.4 SEM-based Validity

This was the main phase of the analysis whereby the ability factors were linked to performance measures (latent variable or observed variable) in a bifactor SEM model. A bifactor approach can estimate the effect of the general factor along with the specific effects of other broad ability factors in the model. In a bifactor model, the ability subtest indicators (e.g., AR, DI, MK) are predicted by both a specific ability factor (e.g., quantitative) and a general factor (e.g., *g*). Hence, the specific ability factors represent variance in the ability tests' scores that are not accounted for by the general factor. Occasionally, the effect of some specific factors on their indicators (i.e., factor loadings) become weak or even negative, indicating that the general factor fully accounts for most of the effect. Substantial differences between subtests' loadings on their respective specific factors and general factor shows how each specific construct is related to the general factor. For example, low (or negative) subtests' loadings on their specific factors, while large on the general factor indicates that the specific construct is fundamental to *g* to the extent that it gets absorbed into it and cannot be separated. Conversely, subtest loadings that are large on their specific factors, while low on general factor indicate that the specific construct diverges from the construct of *g*.

Fitting a bifactor model can be a suitable approach for the current investigations because the relations of both general ability and specific abilities with performance measures are simultaneously estimated. Every ability factor, including the *g* factor, will have a path (i.e., regression) coefficient showing its effect on performance criteria, controlling for other abilities in the model. Thus, the unique contribution of every ability can be estimated.

3.6 Model Fit Indices

All CFA and SEM models were estimated using *maximum likelihood* (ML). Model fit was assessed according to several goodness-of-fit indices, including the Comparative Fit Index (CFI; Bentler, 1995), Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1992), and Standardized Root Mean Square Residual (SRMR; Hu & Bentler, 1999). As suggested by some researchers (Hooper et al., 2008; Hu & Bentler, 1999; MacCallum et al., 1996), good fit of the hypothesized model to the observed data requires a value close to .95 for CFI, although any value over .90 is considered acceptable. Similarly, RMSEA and SRMR values below or near to .06 indicate good fit, although values as high as .08 are considered acceptable. For chi-square (χ^2), due to the large sample used in all primary studies, it was not considered for judging model fit, although it was reported.

3.7 Interpretation of Results

Regarding the interpretation of the resulting effect sizes, the normative correlation guidelines suggested by Gignac and Szodorai (2016) were considered: .10, .20, and .30 indicate relatively small, typical, and relatively large, respectively.

3.8 Statistical Software

All data were analyzed using different R packages (R Core Team, 2019), including *lavaan*, *sem*, *psych, nFactors, and metaSEM.* Some graphs were charted using *lavaanPlot* package. The *metaSEM* package (Cheung, 2015) was used for the meta-analytic preliminary study.

CHAPTER 4: RESULTS

This chapter reports the results yielded from the four studies in this thesis; meta-analytic preliminary study, primary validation study, cross-validation study, and cross-occupation validation study. A brief summary is added at the end of each study and a general summary is presented at the end.

4.1 Preliminary Study

The goal of this study was to examine the factor structure of the AFOQT by means of EFA procedures. Because subsequent studies will make use of the AFOQT subtests for modeling broad ability factors, there was a need to interpret the content of the underlying factors under different assumed structures. To provide dependable results, a meta-analytic SEM was applied on the AFOQT data that were collected from prior investigations. Subtest intercorrelations were meta-analyzed to produce a weighted pooled correlation matrix for further EFA assessment. One- to six-factor EFA models were examined, with the most focus on the five-factor model previously suggested for the AFOQT. The five-factor model was also further examined using second-order EFA to assess whether the five first-order factors remain after accounting for a higher order *g* factor.

4.1.1 Heterogeneity of the Correlation Matrices

The *Q* statistic of the heterogeneity test was significant ($Q_{(693)} = 3476.756$, $p < .05$), indicating that there is substantial heterogeneity in the correlation matrices. Heterogeneity (I^2) statistics of the correlation coefficients (percentages of total variation) are reported in Table 5. The I^2 for the 120 correlation coefficients varied between 43% (GS/TR) and 100% (WK/RC, DI/AR, and RC/VA). This suggests that there is large between-study heterogeneity for most coefficients and that a large part of the variance is at the study level. Heterogeneity tests, therefore, indicates that the assumption that the synthesized correlation matrices are derived from the same population is questionable. A potential cause of such an inflation in the between-studies variance is the small number of correlations included in the meta-analysis (e.g., Sidik and Jonkman, 2007). In any case, the interpretation of the subsequent EFA results should be taken with caution, even though that the heterogeneity of the correlation matrices seemed to have little impact on fit of the structural models (e.g., Jak & Cheung, 2019).

Table 5. *Heterogeneity (I2) Statistics of 120 Correlation Coefficients ((k = 9–11) Between 16 AFOQT Subtests (Above Diagonal) and Weighted Pooled Correlation Matrix for the Relationships Between the Subtests (Below Diagonal)*

	VA	**AR**	**RC**	**DI**	**WK**	**MK**	**MC**	**EM**	**SR**	**IC**	**BC**	**TR**	**AI**	**RB**	**GS**	**HF**
VA	1	62	100	86	99	51	51	46	85	78	74	76	85	48	50	48
AR	.43	1	59	100	50	100	47	46	99	80	92	88	90	66	84	60
RC	.56	.43	1	54	100	52	49	45	48	82	64	44	97	45	99	44
DI	.39	.58	.42	1	70	99	47	45	88	81	66	72	79	46	47	45
WK	.50	.35	.66*	.36	1	48	78	46	49	72	95	46	93	48	63	48
MK	.41	.60	.39	.45	.31	1	46	46	73	81	87	62	97	56	49	46
MC	.32	.34	.33	.31	.24	.32	1	51	68	66	84	96	93	72	99	88
EM	.13	.23	.12	.24	.08^{ns}	.26	.34	1	56	82	78	46	95	50	45	49
SR	.30	.47	.28	.49	.23	.43	.27	.29	1	80	96	88	94	56	57	76
IC	.19	.23	.17	.29	.16	.16	.32	.27	.30	1	80	50	91	46	59	45
BC	.24	.31	.20	.33	.17	.32	.33	.37	.37	.29	1	51	93	89	61	77
TR	.19	.27	.18	.32	.12	.28	.16	.19	.36	.20	.40	1	95	46	43	63
AI	.11	.12	.14*	.17	.18	-.02^{ns}	.31	.08^{+}	.12**	.44	.05	.03	1	90	92	87
RB	.26	.28	.19	.27	.16	.31	.39	.32	.32	.31	.41^{ns}	.19^{ns}	.12	1	47	50
GS	.41	.36	.46	.30	.44	.40	.45	.21	.22	.25	.19	.09	.27	.28	1	46
HF	.24	.22	.19	.23	.16	.26	.29	.25	.30	.22	.34	.23	.07*	.33	0.20	1

Note. Gray color is Heterogeneity *(I^2)* Statistics indicating the percentage of the variation; the Correlations without significance signs were all significant at $p < .001$.

4.1.2 Pooled Correlation Matrix

The correlations between subtests were pooled using a multiple-group CFA model (Cheung & Chan, 2009). Table 5 also shows the pooled intercorrelation matrix among AFOQT subtests. The majority of correlations were significantly larger than zero ($p < .001$; $M = .28$, $SD = .12$), perhaps due to the large combined sample size aggregated from the 11 matrices ($N = 21{,}102$). Out of 120 correlation coefficients, only five were not significant (WK/EM, MK/AI, EM/AI, BC/RB, and TR/RB; $p > .05$). The magnitude of coefficients was distributed as follows: 7 were below .10; 25 were between .10 and .19; 34 were between .20 and .29; 33 were between .30 and .39; 16 were between .40 and .49; and 5 were .50 or above. In spite of the high range restriction of military pilot samples, the pooled correlations showed some high associations between subtests such as that between the Word Knowledge and Reading Comprehension tests ($r = .66$).

4.1.3 Factor Extraction Criteria Comparisons

The pooled correlation matrix was then used to perform EFA in order to assess the AFOQT's internal structure based on one- to six-factor models. The minimum average partial method (MAP) suggested two factors; eigenvalue > 1 suggested four factors, with the largest four eigenvalues ranging from 2.45 to 1.16, exploratory graph analysis (EGA) also suggested four factors (Figure 2b); scree plots and Horn's parallel analysis (HPA) suggested five factors (Figure 2a), and BIC and SSBIC suggested eight factors (Decreased respectively from the single-factor through the eight-factor as follows: 32423/32753; 17352/17635; 9536/ 9775; 4891/ 5088; 1104/ 1263; 675/799; 226/318; 62/126).

a b

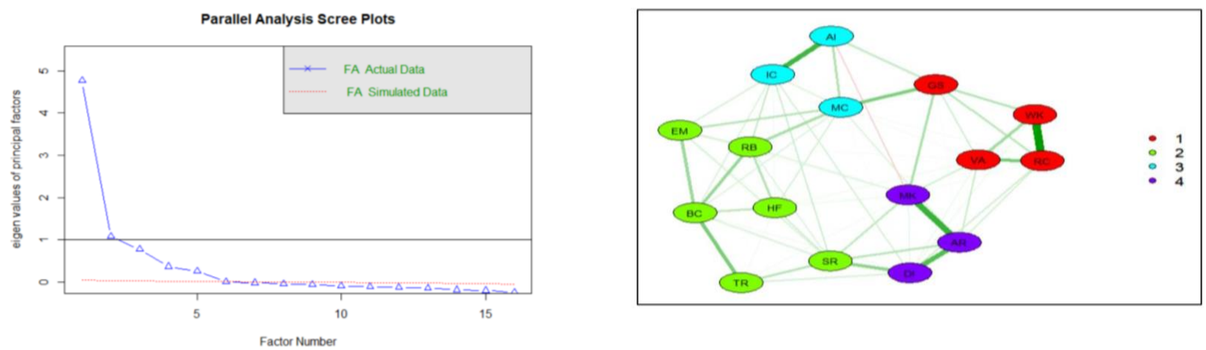


Figure 4. Factors Extracted based on (a) Parallel Analysis; (b) Exploratory Graph Analysis

AFOQT researchers recommended five factors. The different factor solutions suggested by different criteria justify examining the AFOQT factor structure more expansively. Accordingly, models with one, two, three, four, five, and six factors were examined for adequacy. Some argue that over-extraction of factors in EFA may be preferred to under extraction because loading estimates of true factors tends to include less error in the case of overextraction (Wood et al., 1996).

4.1.4 EFA for the Five-Factor Model

Table 7 showed the results of the EFA and second-order EFA for the five-factor AFOQT model. The first-order five-factor solution showed somewhat consistent results to the theorized model suggested for the AFOQT. The extracted factors consist of a verbal ability factor, a spatial ability factor, a quantitative ability factor, an aviation-related acquired knowledge factor, and a perceptual speed factor. However, two notable differences are particularly worth mentioning. First, the General Science test had its primary loading on the verbal ability factor (.35) and secondary loading on the spatial ability factor (.28), but had no significant loading on the acquired aviation knowledge factor (.10), although this factor has been suggested as a primary factor for the GS (Carretta & Ree, 1996). Second, the loading of the Scale Reading test on the quantitative ability factor (.42) was higher than its loading on the theorized factor of perceptual speed (.35), although both were significantly greater than zero. The Block Counting test had an even higher loading on perceptual speed than Scale Reading did (.38). The rest of the loadings and cross-loadings matched the suggested five-factor model to a reasonable degree.

Table 6. *First-order and Second-order Exploratory Factor Analysis of the 16 AFOQT Subtests*

	EFA					Second-order EFA						
	Verbal	Spatial	Quantitative	Knowledge	Perceptual	*g*	Verbal	Spatial	Quantitative	Knowledge	Perceptual	
VA	.57					.53	.42					**VA**
AR			.89			.71			.42			**AR**
RC	.88					.56	.63					**RC**
DI			.55		.28	.63			.28			**DI**
WK	.90					.47	.63					**WK**
MK			.71			.64			.33			**MK**
MC		.56				.47		.29		.21	-.36	**MC**
EM		.55				.33		.39				**EM**
SR			.42		.35	.55		.24	.23			**SR**
IC		.23		.50		.35		.27		.45		**IC**
BC		.61			.38	.41		.59				**BC**
TR					.45	.34		.33			.27	**TR**
AI				.89		.22				.81		**AI**
RB		.64				.40		.44			-.20	**RB**
GS	.35	.28		**.10**		.50	.27				-.37	**GS**
HF		.48				.33		.39				**HF**

Note. VA = Verbal Analogies; AR = Arithmetic Reasoning; RC = Reading Comprehension; DI = Data Interpretation; WK = Word Knowledge; MK = Math Knowledge; MC = Mechanical Comprehension; EM = Electrical Maze; SR = Scale Reading; IC = Instrument Comprehension; BC = Block Counting; TR = Table Reading; AI = Aviation Information; RB = Rotated Blocks; GS = General Science; HF = Hidden Figures

4.1.5 Other Factor Solutions

Table 8 presents the results of extracting one to six AFOQT factors with the *Promax* rotation. The table includes information about factors' loadings, factors intercorrelations, factors' eigenvalues, the proportion of variance explained by each factor, substantive interpretability of each factor solution, and model fit indices (TLI, RMSEA, and BIC). The five- and six-factor models showed similar fit statistics, with comparable TLIs (.97) and RMSRs (.04), but a smaller BIC value than for the six-factor solution (ΔBIC = 375). However, one of the factors in a six-factor solution had only one salient subtest (factor 5), which may be an indication for factor overextraction. The four-factor solution also showed an acceptable model fit, but substantial differences (ΔTLI = .06, ΔRMSR = .03, ΔBIC = 3791), suggested superiority of the five-factor model over the four-factor model. The remaining one-, two, and three-factor solutions yielded poor model fit statistics. Thus, the five-factor model remained the most plausible solution for the AFOQT meta-analytic data.

Four important points might be of interest to note from the EFA results. First, the four-factor model appeared a competitive structure for the AFOQT because it met the requirement of simple structure, with each factor marked by two to five salient loadings, and none of the tests had salient cross-loadings with other factors (≥ .30). According to this solution, subtests of quantitative (Arithmetic Reasoning, Data Interpretation, Math Knowledge) and perceptual speed (Table Reading, Scale Reading) factors may have a common variance that can be accounted for by one broad factor (i.e., quantitative-perceptual ability). The other three factors are identical to those that emerged in the five-factor solution (verbal, spatial, and aviation-related acquired knowledge).

Table 7. *AFOQT Exploratory Factor Analysis: One to Six Oblique Factor Solution for the Pooled Correlation Matrix*

	1-factor	2-factor		3-factor			4-factor				5-factor					6-factor					
	g	V+Q	S+P	Q+S	V+Q	K+M	V	S	Q+P	K	V	S	Q	K	P	Q	S	V	K	??	M
VA	0.61	0.69			0.64		0.61				0.57							0.53			
AR	0.69	0.43	0.32	0.44	0.43		0.26		0.63				0.89			0.90					
RC	0.62	0.94			0.87		0.83				0.88							0.81			
DI	0.67	0.37	0.36	0.43	0.35				0.69				0.55		0.28	0.64					
WK	0.53	0.88			0.81		0.77				0.90							0.85			
MK	0.65	0.37	0.34	0.50	0.38		0.29		0.42				0.71			0.64					0.28
MC	0.58		0.43	0.26		0.39	0.25	0.52				0.56					0.41				0.36
EM	0.41		0.61	0.52				0.56				0.55					0.56				
SR	0.60		0.57	0.62					0.63				0.42		0.35	0.51	0.22				
IC	0.44		0.49	0.27		0.52		0.25		0.45		0.23		0.50			0.29		0.46		
BC	0.53		0.73	0.71				0.59	0.21			0.61			0.38		0.63			0.21	
TR	0.40		0.48	0.55					0.45						0.45					0.76	
AI	0.26	**.14**	**.14**			0.67				0.93				0.89					0.90		
RB	0.51		0.60	0.50				0.65				0.64					0.71				
GS	0.57	0.56			0.52	0.30	0.58	0.23			0.35	0.28						0.26			0.50
HF	0.43		0.49	0.45				0.48				0.48					0.51				
EV	4.69	2.97	2.9	2.87	2.76	1.18	2.45	1.89	2.04	1.16	2.10	1.92	1.94	1.13	0.89	2.01	1.85	1.89	1.12	0.67	0.75
PE		50%	50%	42%	41%	17%	32%	25%	27%	15%	26%	24%	24%	14%	11%	24%	22%	23%	13%	8%	9%
		1		1			1				1					1					
		.63		0.55	1		0.45	1			0.49	1				0.65	1				
				0.40	0.30	1	0.51	0.63	1		0.67	0.65	1			0.61	0.41	1			
							0.31	0.39	0.16	1	0.35	0.47	0.36	1		0.26	0.41	0.27	1		
											0.06	0.29	0.29	0.06	1	0.55	0.56	0.31	0.12	1	
																0.29	0.29	0.4	0.15	-0.14	1
TLI	.65	.78		.85			.91				.97					.97					
RMSR	.12	.10		.08			.07				.04					.04					
BIC	32423	17354		9537			4896				1105					730					

Note. numbers written in grey are those with loadings between .20 to .30. *g* = general factor; V = Verbal; Q = Quantitative, S = Spatial; P = Perceptual Speed; K = Acquired Job Knowledge; M = Mechanical; EV = Eigenvalue; PE = Proportion of variance explained; VA = Verbal Analogies; AR = Arithmetic Reasoning; RC = Reading Comprehension; DI = Data Interpretation; WK = Word Knowledge; MK = Math Knowledge; MC = Mechanical Comprehension; EM = Electrical Maze; SR = Scale Reading; IC = Instrument Comprehension; BC = Block Counting; TR = Table Reading; AI = Aviation Information; RB = Rotated Blocks; GS = General Science; HF = Hidden Figures

Second, forcing the six-factor model resulted in the emergence of a specific factor composed of Mechanical Comprehension and General Science subtests, which might be indicative of mechanical knowledge factor. This finding exposes that the Mechanical Comprehension test has content that differs from the four spatially-oriented tests in the AFOQT (Block Counting, Rotated Blocks, Electrical Maze, Hidden Figures). Another sign of the divergence of this test from other spatial ability tests is the pattern of loadings in the three-factor model, where all spatial ability tests loaded with the quantitative ability subtests, while the MC test loaded on the factor indicative of acquired aviation knowledge. Third, the Instrument Comprehension test showed cross-loadings with spatial ability factors across all models ranging from .23 (five-factor model) to .29 (six-factor model). This suggests that Instrument Comprehension subtest is loaded on spatial ability construct, in addition to the theorized aircrew aptitude construct, which may also explain its robustness in predicting pilot performance. The content of this test, which measures the ability to determine aircraft attitude from illustrations of flight instruments, actually suggests such a shared variance. Fourth, the Aviation Information test failed to achieve a salient pattern coefficient on any factor when one-factor and two-factor models were forced, which indicates its nature as a domain-specific test that is less loaded with general ability. Regarding factors intercorrelations, the strongest were quantitative ability with verbal ability and spatial ability, whereas the weakest were aviation-related acquired knowledge with other factors, particularly perceptual speed.

4.1.6 Hierarchical EFA for the Five-Factor Model

Given the plausibility of the five-factor EFA solution revealed by the results, it was accordingly transformed to a second-order model with the *Schmid-Leiman orthogonalization*. The results are also presented in Table 6. After transformation, specific factors of verbal ability, quantitative ability, spatial ability, and acquired job knowledge were almost identical to those in the EFA solution in terms of the primary subtests' composition, although their loadings degraded substantially. One major difference between the two models was that the factor of perceptual speed didn't hold in the second-order model as a distinct factor, and both of its indicators (Scale Reading and Table Reading) grouped with spatial ability indicators. This seems reasonable since spatial and perceptual speed abilities shared considerable common variance, and there was a large ratio of factors to observed scores. In addition to its loading on spatial ability, the Table Reading test had another sizable loading (.27) on a fifth factor, accompanied by three subtests (General Science, Mechanical Comprehension, Rotated Block) that were negatively loaded on the same factor.

The hierarchical *g*–factor accounted for 47% of the common variance, with loadings ranging from .22 (Aviation Information) to .71 (Arithmetic Reasoning). Because the Aviation Information subtest is a specific-domain knowledge test, the pattern of its loadings on its respective factor (.81) was notably larger than its loading on the *g* factor (.22). A reversed pattern of factor loadings was noted for the three quantitative ability subtests (Arithmetic Reasoning, Data Information, and Math Knowledge), where all had a much larger loading on the *g* factor than on their respective factor ($\Delta\ \lambda \geq .29$). An interesting pattern was observed for the subtests of Mechanical Comprehension, Scale Reading, Instrument Comprehension, and Table Reading, where they had two noteworthy loadings, although some were below .30, in addition to their primary significant loadings. This pattern reveals important information about the constructs underlying these subtests. For example, the Instrument Comprehension subtest is often presented as a domain-specific subtest related to aviation, which is supported by both the lower-order and second-order EFAs in this study. However, the noteworthy loadings it also showed on the *g* factor (.35) and spatial ability (.27) suggesting that the construct underlying this subtest should be interpreted more broadly. Overall, the hierarchical EFA appeared to favor a four first-order and one higher-order (g) factor model for the 16 AFOQT subtests.

4.1.7 Summary

This study was useful in understanding the underlying constructs of the AFOQT subtests. Results will inform subsequent studies and facilitate the selection of subtests representing particular broad ability constructs. Traditionally, AFOQT researchers had advocated a five-factor model for the AFOQT, even when the subtests were reduced in number from 16 to 11 (Form S), and then to ten (Form T). This study supported the adequacy of five latent ability factors explaining the AFOQT data. However, the four-factor model (verbal, spatial, quantitative/perceptual, and acquired aviation knowledge) seemed another plausible factor model for the AFOQT, with even better characteristics in terms of simple structure and its hierarchical structure.

4.2 Primary Validation Study

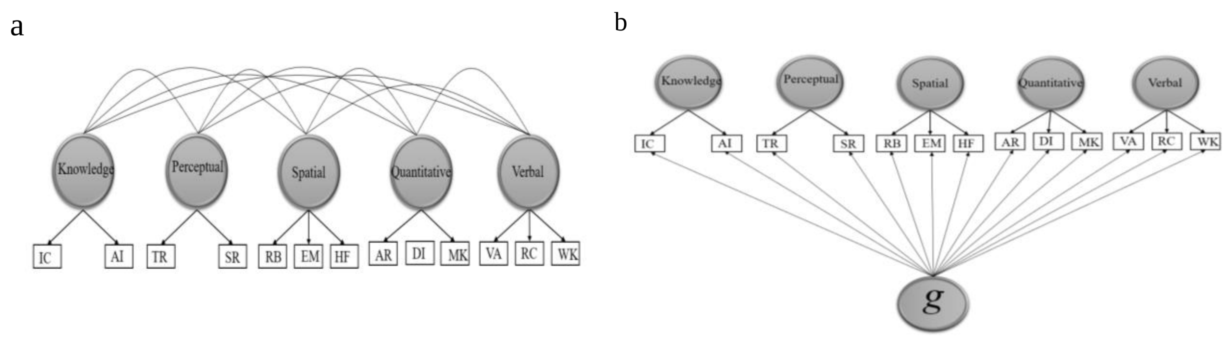
The goal of this study was to demonstrate the validity of five particular cognitive abilities for predicting pilot performance. A predictive model based on AFOQT five-factor model was introduced to serve as a basis for several succeeding investigations attempting to replicate the findings. The relative contributions of the general and specific factors were examined via bifactor modeling approach. The data of this study were obtained from Carretta and Ree's (1995) primary study, which included an intercorrelation matrix for the 16 AFOQT subtests and five performance measures for 7,563 USAF pilot students.

4.2.1 Modeling Performance

From the five performance measures reported in Carretta and Ree's (1995) study, two latent factors were extracted to represent pilots' performance at the primary (two measures) and advanced (two measures) phases of training. Both latent factors were constructed from daily flight and check flight ratings collected at each phase of training. The fifth measure, academic performance, was kept as observed. Hence, two latent factors and one observed score were used as dependent variables in this study.

4.2.2 Modeling Cognitive Abilities

Prior to estimating the relations between cognitive abilities and flight performance criteria, I conducted a CFA to assess the fit of the suggested five-factor model. Because the spatial ability construct in the long AFOQT is manifested by five indicators (Mechanical Comprehension, Rotated Block, Electrical Maze, Block Counting, and Hidden Figure), I excluded two subtests (Mechanical Comprehension and Block Counting) that were found strongly loaded on other ability factors, in addition to their primary loadings on spatial ability factors (see preliminary study). I also excluded the General Science subtest that is often posited as an indicator for the acquired knowledge factor, together with the Instrument Comprehension and Aviation Instrument subtests. This subtest did not show a significant loading on its theorized underlying factor (see preliminary study) and also contains general science content that is not specific to aviation science, which may diverge from the aviation-related acquired knowledge construct that is purported to be measured by this factor. Accordingly, each ability factor was indicated by two (perceptual speed and aviation-related acquired knowledge) to three (verbal, quantitative, and spatial abilities) indicators. The five AFOQT factors were modeled by correlated-factor and bifactor models because both will assist in the validation process. Figures 3a and 3b display the two models used in this study.



Note. VA = Verbal Analogies; RC = Reading Comprehension; WK = Word Knowledge; AR = Arithmetic Reasoning; DI = Data Interpretation; MK = Math Knowledge; RB = Rotated Blocks; EM = Electrical Maze; HF = Hidden Figures; TR = Table Reading; SR = Scale Reading; IR = Instrument Comprehension; AI = Aviation Information.

Figure 5. Five Ability Factors. (a) Correlated-factor Model; (b) Bifactor Model.

Both models fit the data similarly well: $\chi^2(57) = 1995.63$, $p \leq .001$; CFI = .92; RMSEA = .07; SRMR = .05 for the correlated-factor model and $\chi^2(54) = 2039.61$, $p \leq .001$; CFI = .92; RMSEA = .07; SRMR = .05 for bifactor model. Table 9 presents factor loadings for both models. Concerning the correlated-factor model, subtests' loadings on their respective factors were all significant ($p \leq .001$), ranging from .48 (HF on spatial ability factor) to .79 (VA on Quantitative ability factor). All inter-factor correlations were significant ($p \leq .001$) and ranged from 0.20 (between quantitative ability and aviation-related acquired knowledge) to .80 (between quantitative ability and perceptual speed).

Table 8. *Factor Loadings of Ability Factors Based on Correlated-factor and Bifactor Models (Primary Study)*

Model	Factor	Verbal			Quantitative			Spatial			Perceptual		Knowledge	
		VA	RC	WK	AR	DI	MK	RB	EM	HF	TR	SR	IC	AI
Correlated-factor	Specific	.79	.71	.68	.78	.67	.69	.60	.49	.48	.55	.67	.69	.59
	General	-	-	-	-	-	-	-	-	-	-	-	-	-
Bifactor	Specific	.66	.53	.47	-.10*	.14**	.64*	.51	.37	.35	.36	.40	.64	.58
	General	.49	.42	.50	.72	.74	.65	.37	.30	.31	.34	.61	.27	.10

Note. All factor loadings were significant at $p < .001$, except those indicated in the table that were significant at a higher level.
* $p < .01$. ** $p < .01$.

Factor loadings in the bifactor model were all significant. The differences in factor loadings between those of general ability and those of specific factors were not large, with exception of two subtests in the quantitative factor (AR and DI) that showed much higher loadings on the general factor, and the two subtests of aviation-related knowledge factor that showed much higher loadings on their specific factor. Accordingly, the findings indicated that both models are viable representations for the AFOQT data.

4.2.3 Correlation-based Validity

Table 10 displays the correlations between the five cognitive abilities and the three flight performance criteria. As shown in the table, the associations with flight performance seemed stronger for aviation-related acquired knowledge, perceptual speed, and quantitative ability than for verbal and spatial abilities. The verbal ability factor had no association with advanced flight performance.

Table 9. *Correlations Between Ability Factors and Pilot Performance Criteria (Primary Study)*

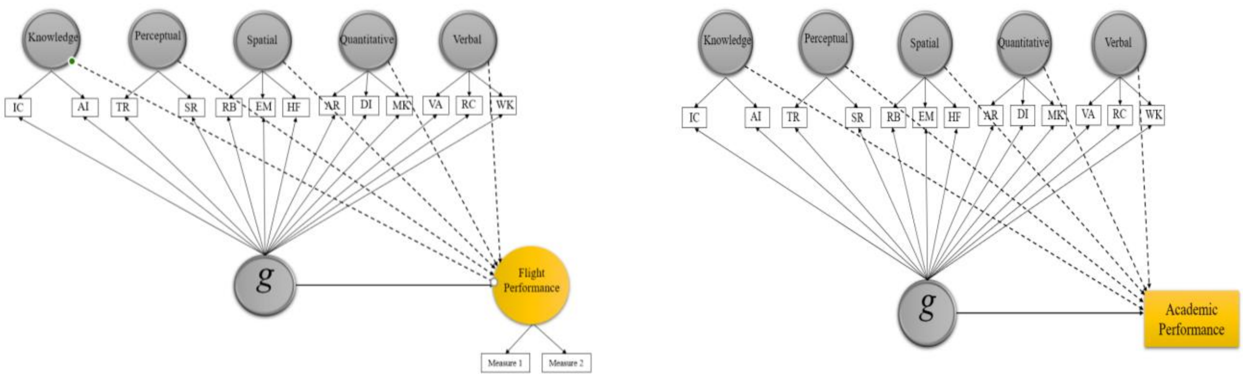
Ability Factor	Primary	Advanced	Academic
Verbal Ability	0.06^{***}	-0.02^{ns}	0.19^{***}
Quantitative Ability	0.20^{***}	0.10^{***}	0.26^{***}
Spatial Ability	0.15^{***}	0.06^{**}	0.12^{***}
Perceptual Speed	0.37^{***}	0.20^{***}	0.23^{***}
Acquired Knowledge	0.49^{***}	0.16^{***}	0.13^{***}

Note. gray color indicates negative estimate or nonsignificant positive estimate (p < .05). Model fit: $\chi^2(110) = 2761.44$, $p \leq .001$; CFI = .92; RMSEA = .06; SRMR = .04.
^{ns} $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

4.2.4 SEM-based Validity (Bifactor Predictive Model)

Figures 4a displays the structural models used to estimate the predictive relations between cognitive abilities and pilot performance. In this figure, the effects of the five ability factors (or their residuals) on each latent/observed performance measure were separately estimated. Similar to multiple regression, each effect represents the degree to which an ability factor uniquely predicts performance criterion after accounting for the effects of the other ability factors in the model (and after partialling out performance variance that is domain-general).

a b



Note. Solid lines represent paths from the general ability factor; dashed lines represent paths from the specific ability factors. VA = Verbal Analogies; RC = Reading Comprehension; WK = Word Knowledge; AR = Arithmetic Reasoning; DI = Data Interpretation; MK = Math Knowledge; RB = Rotated Blocks; EM = Electrical Maze; HF = Hidden Figures; TR = Table Reading; SR = Scale Reading; IR = Instrument Comprehension; AI = Aviation Information.

Figure 6. Bifactor SEM Model on (a) Latent Flight Performance; (b) Observed Academic Performance

Table 11 displays the standardized path coefficients for the relations. For the primary phase of training, aviation-related acquired knowledge (β = .43), general factor (β = .26), and perceptual speed (β = .23) were the three positive and significant predictors ($p < .001$). Fairly similar results about the relative roles of abilities were found for the advanced phase of training, with perceptual speed (β = .16), general factor (β = .12), and aviation-related acquired knowledge (β = .12) being the only positive and significant predictors in the model ($p < .001$). Verbal, quantitative, and spatial abilities did not contribute positively in the predictive model; in fact, the unique effects of verbal ability were consistently negative. The academic performance outcome showed a different pattern, where four of the five specific abilities had significant effects (β = .04–.08, $p < .01$) above the effect of general ability. Additionally, general ability had a larger effect on academic performance than any of the specific abilities (β = .24).

Table 10. *Prediction of Pilot Performance by General Ability and Specific Abilities Via Bifactor Models (Primary Study)*

	Primary	Advanced	Academic
Verbal Ability	-.14***	-.10***	.04**
Quantitative Ability	-.07*	-.02^{ns}	.08***
Spatial Ability	-.09***	-.03^{ns}	-.06***
Perceptual Speed	.23***	.16***	.06**
Acquired Knowledge	.43***	.12***	.08***
g Bifactor	.26***	.12***	.24***

Note. Models fit were as follows. Primary: $\chi^2(73) = 2246.43$, $p < .001$; CFI = .92; RMSEA = .06; SRMR = .05; Advance: $\chi^2(73) = 2104.83$, $p < .001$; CFI = .93; RMSEA = .06; SRMR = .04; Academic: $\chi^2(61) = 2109.58$, $p < .001$; CFI = .92; RMSEA = .07; SRMR = .05.
ns $p > .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

4.2.5 Summary

In this primary study, the predictive relations between cognitive abilities and pilot performance were assessed through bifactor predictive SEM models. Among the six abilities assessed, only two specific abilities, along with the general ability, showed a unique contribution to the prediction of three performance measures. This is different from the result obtained when associating predictors and criteria in a correlated-factor model, where almost all specific abilities demonstrated significant, positive relations with pilots' subsequent training performance. Isolating domain-general from domain-specific effects showed negligible domain-specific effects for verbal, quantitative, and spatial abilities on the primary and advanced training performance criteria. Aviation acquired knowledge, general ability, and perceptual speed were the three abilities most contributing to performance in these two phases of training. For academic achievement, although this criterion was significantly predicted by four of the five domain-specific factors (spatial ability was the exception), all relations were of small magnitude (β = .04 to .08). General ability contributed a little more than the specific abilities to the prediction of academic performance (β = .24), but not so for the actual flying performance. The few relations between specific abilities and performance criteria that remained significant after removing the general-domain effect indicates that the contribution of some specific cognitive abilities in flight performance is more than only *g*.

4.3 Cross-validation Study

The objective of this study was to evaluate the consistency of the predictive model across samples and criteria. Three pilots' samples were used, all featuring multiple flight performance criterion measures. A similar procedure to that applied in the validation study was used in this cross-validation. First, pilot performance measures were identified. Second, CFA measurement models were tested to assess how well correlated-factor and bifactor models describe the AFOQT data. Third, correlations between the ability factors and performance criteria were assessed to obtain an initial idea about the expected relations among variables in the models. Fourth, a bifactor modeling approach was applied to examine the effect of general and specific ability factors on the performance criteria.

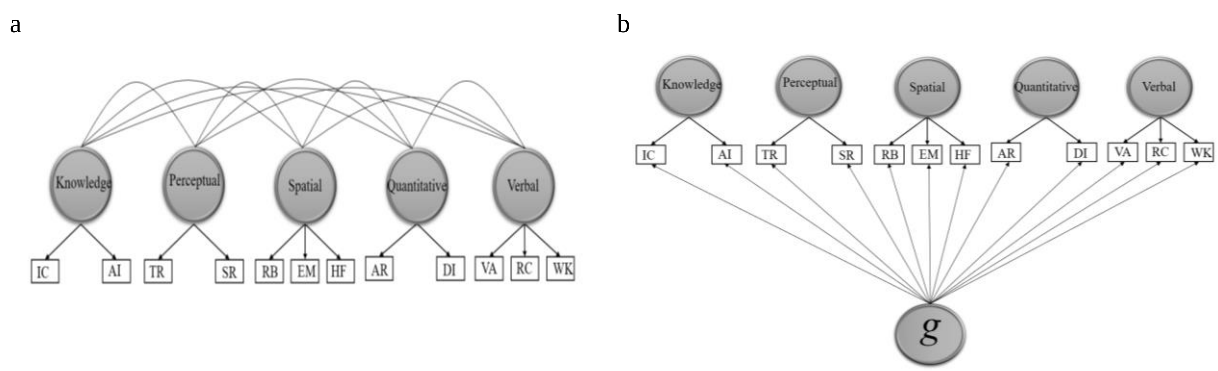
4.3.1 Modeling Performance

Data of the three pilots' samples were obtained from the primary studies of Duke and Ree (1996; Sample 1), Olea and Ree (1994; Sample 2), and Carretta and Ree (1997; Sample 3), which all reported intercorrelation matrices among AFOQT subtest scores and multiple flight performance measures. Sample 1 reported three measures, all of which were modeled as observed scores because each represented a distinct type of performance. These are: (1) the exceedance of average flying hours in the primary phase of training, (2) the exceedance of average flying hours in the advanced phase of training, and (3) the students' rank order upon completion of the flight training course. Sample 2 reported six performance measures (see Table 4 of Chapter 3), three of which were used in this study, while the remaining three were used in the cross-occupation validation study. A consideration taken into account when modeling measures for each study was that they covered different scopes (i.e., general score, specific score) and phases of training (i.e., primary, advanced). While the measures utilized in the cross-occupation study were modeled as indicators of a latent variable, all three measures in this study were retained as observed scores, including (1) check flight ratings from the primary phase of training, (2) check flight ratings from the advanced phase of training, and (3) an overall composite of the five performance measures reported in the study. Sample 3 had the largest number of performance measures, six check flight ratings and 11 academic grades. Check flight ratings during the primary (three measures) and advanced (three measures) training phases were modeled as two latent factors. The remaining 11 measures were end-of-course grades in early (four measures), middle (four measures), and late (three measures) stages of training. A separate EFA for these three sets of academic performance measures indicated that they are best explained by one performance factor and thus, they were modeled as one latent factor for overall academic grade. This factor had an eigenvalue of 1.25 explaining 12% of the variance in subtest scores, with subtests' loadings ranging from .21 to .54. Accordingly, the ability factors in this study were examined for their effects on two latent factors of flight performance in two phases of training, as well as on one latent factor representing academic performance.

4.3.2 Modeling Cognitive Abilities

Regarding modeling cognitive abilities, two slight differences (Samples 1 and 2) and one major difference (Sample 3) can be highlighted between the CFA models (and the subsequent SEM models) tested in the primary validation study, and those tested in this cross-validation study. A slight modification had to be made to the CFA models of the Sample 1 and 2 data for the specification of the quantitative ability construct. For both datasets, the Math Knowledge subtest

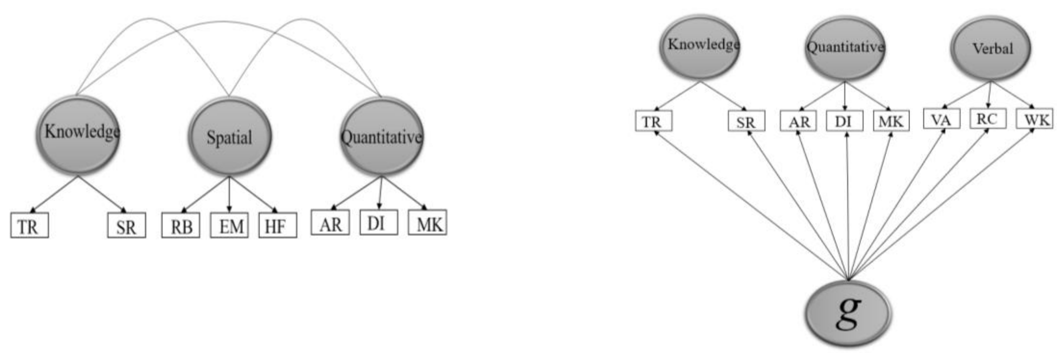
(MK) was omitted, so the quantitative ability construct was represented by the remaining two subtests (Arithmetic Reasoning and Data Interpretation). Redefining the construct in this cross-validation study enabled the models to be identified. Also conceptually, the MK subtest seemed the least contributing subtest to the quantitative factor, and the *g* factor, among quantitative subtests. CHC-based theoretical classification of AFOQT subtests (see Table 2 of Chapter 2) suggests some differences between the MK subtest and the other indicators of quantitative factor (AR and DI). MK is categorized as Quantitative Acquired Knowledge, while AR and DI are categorized as Quantitative Reasoning related to Fluid Intelligence. A representation of this modified version of the five-factor model is portrayed in Figures 5a and 5b.



Note. VA = Verbal Analogies; RC = Reading Comprehension; WK = Word Knowledge; AR = Arithmetic Reasoning; DI = Data Interpretation; MK = Math Knowledge; RB = Rotated Blocks; EM = Electrical Maze; HF = Hidden Figures; TR = Table Reading; SR = Scale Reading; IR = Instrument Comprehension; AI = Aviation Information.

Figure 7. Five Factor model (a) Correlated-factor model; (b) Bifactor model.

a b



Note. VA = Verbal Analogies; RC = Reading Comprehension; WK = Word Knowledge; AR = Arithmetic Reasoning; DI = Data Interpretation; MK = Math Knowledge; RB = Rotated Blocks; EM = Electrical Maze; HF = Hidden Figures; TR = Table Reading; SR = Scale Reading; IR = Instrument Comprehension; AI = Aviation Information.

Figure 8. Three Factor model (a) Correlated-factor model; (b) bifactor model.

For Sample 3, because the reproduced intercorrelation matrix included only nine of the 16 AFOQT subtests, it was not possible to extract all targeted five factors. Only verbal ability (3 indicators), quantitative ability (3 indicators), and aviation-related acquired knowledge (2 indicators) were potentially represented in this data. However, even with this limitation, this data still seemed useful for cross-validation, particularly due to the large number of performance criteria reported in the study (17), which would add value to the current investigation. The newly introduced three-factor model are portrayed in Figures 6a and 6b.

Results for the CFA measurement models of the three samples are reported in the extended results reported in Appendix B. Both correlated-factor and bifactor models adequately described the data for all three samples. Factor loadings of both models were all significantly greater than zero at $p \leq .001$, with only a few exceptions. The inter-factor correlations in the correlated-factor models were all significant ($p \leq .001$). Table 12 presents the standardized loading coefficients of the subtests on both *g* factor and specific ability factors based on bifactor models. Overall, subtests of aviation-related acquired knowledge (Aviation Information and Instrument Comprehension), as well as the Word Knowledge subtest of verbal ability had loadings on their respective factors notably higher than their loadings on the *g* factor. This set of subtest seems the least related to the general factor. Conversely, subtests of quantitative abilities (Data Interpretation, Arithmetic Reasoning, Math Knowledge), as well as the Scale Reading subtest of perceptual speed had loadings on the *g* factor notably higher than that loadings on their respective factors. This set of subtest seems the most related to the general factor. The remaining subtests showed fairly balanced loadings between their own factors and *g* factor. There were some differences between the three cross-validation samples and the primary validation sample in the pattern of subtests' loadings on general and specific factors. The largest variation across the four samples was that noted for the subtests of AR (-.07–.82), VA (-.17–.37), and WK (-.35–.03), while the least variation was that noted between HF (-.04–-.01), EM (-.08–0), and RB (-.14–-.05). On the whole, it was concluded that both the three-factor and adapted five-factor models are viable representations of the AFOQT data across Sample 3, and Samples 1 and 2, respectively.

Table 11. *Factor Loadings of Specific and General Factors Based on Bifactor Models (Cross-validation Study)*

		Verbal			Quantitative			Spatial			Perceptual		Knowledge	
		VA	RC	WK	AR	DI	MK	RB	EM	HF	TR	SR	IC	AI
Sample 1	Specific	.48	.63	.71	.35	.35	-	.45	.38	.30	.44	.44	.55	.55
	General	.52	.53	.44	.54	.69	-	.40	.30	.29	.35	.50	.36	.22
Sample 2	Specific	.55	.67	.74	$.06^{ns}$	.28	-	.48	.35	.34	.31	.36	.57	.63
	General	.46	.50	.39	.74	.72	-	.39	.35	.33	.41	.68	.36	.12
Sample 3	Specific	.30	.50	.60	.77	.17	.25	-	-	-	-	-	.58	.57
	General	.67	.62	.62	.70	.72	.54	-	-	-	-	-	.44	.27

Note. All loadings were significant at $p < .001$, except for the AR subtest on quantitative factor in Sample 2
^{ns} $p < .10$.

4.3.3 Correlation-based Validity

Table 13 displays the correlations between the cognitive abilities and nine flight performance criteria (three per sample). Overall, the associations with flight performances seemed stronger for aviation-related acquired knowledge, perceptual speed, and quantitative ability than for verbal and spatial abilities. The verbal ability factor associated weakly with most criteria except with academic achievement (Sample 3), where it was the second most-highly correlated factor ($r = .35$, $p < .001$), after quantitative ability ($r = .46$, $p < .001$). This pattern of cognitive ability-pilot performance association replicated that noted in the primary validation study.

Table 12. *Correlations Between Cognitive Abilities and Pilot Performance (Cross-validation Study)*

	Sample 1			Sample 2			Sample 3		
	Primary	Advanced	Rank	Primary	Advanced	Overall	Primary	Advanced	Academic
Verbal	-0.02^{ns}	0.01^{ns}	0.04^{ns}	$.04^{ns}$	$.00^{ns}$	$.05^{+}$	$.07^{**}$	$.05^{ns}$	$.35^{***}$
Quantitative	0.13^{***}	0.10^{*}	0.20^{***}	$.19^{***}$	$.09^{***}$	$.16^{***}$	$.21^{***}$	$.21^{***}$	$.46^{***}$
Spatial	0.08^{+}	0.12^{**}	0.15^{***}	$.14^{***}$	$.12^{***}$	$.17^{***}$	-	-	-
Perceptual	0.18^{***}	0.17^{***}	0.27^{***}	$.26^{***}$	$.19^{***}$	$.16^{***}$	-	-	-
Knowledge	0.22^{***}	0.22^{***}	0.30^{***}	$.40^{***}$	$.20^{***}$	$.26^{***}$	$.21^{***}$	$.21^{***}$	$.26^{***}$

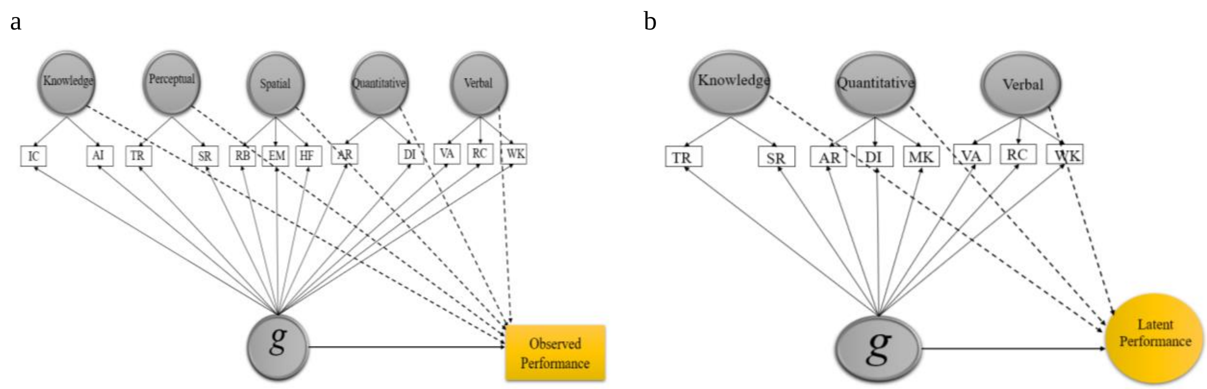
Note. gray color indicates negative estimate or nonsignificant positive estimate (p < .05). Models fir was as follows. Sample 1: $\chi^2(69) = 328.69$, $p < .001$; CFI = .93; RMSEA = .06; SRMR = .04; Sample 2: $\chi^2(50) = 365.11$, $p < .001$; CFI = .95; RMSEA = .03; SRMR = .04; Sample 2: $\chi^2(261) = 1203.10$, $p < .001$; CFI = .94; RMSEA = .06; SRMR = .03.
ns $p > .10$. + $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Across the three studies (see Table 13), among the specific abilities, aviation-related acquired knowledge had the largest correlations with performance measures ($r = .21–40$), except with academic performance in Sample 3 for which it was the weakest predictor ($r = .26$). Additionally, the significance level of its associations never exceeded .001, suggesting strong correlation estimates for this factor. The perceptual speed factor ranked next in association strength with the flight performance measures, with correlations ranging from .17 to .27 (all $p < .001$). Quantitative ability showed only slight superiority over spatial ability in the associations with performance measures, but also had the largest effect size obtained in this study ($r = .46$ with academic performance, $p < .001$; Sample 3). According to this initial analysis, aviation-related acquired knowledge, perceptual speed, quantitative ability, and spatial ability may show noteworthy predictive relations with pilot performance measures, but not so for verbal ability.

However, taking into account the variance attributed to general ability in the next step may change the inference about the relative role of these abilities in flying.

4.3.4 SEM-based Validity (Bifactor Predictive Model)

Figures 7a displays the structural model used for Samples 1 and 2, which is almost identical to that used in the original validation study, and Figure 7b displays the structural model used for Sample 3, which had only three specific ability factors.



Note. Solid lines represent paths from the general ability factor; dashed lines represent paths from the specific ability factors. VA = Verbal Analogies; RC = Reading Comprehension; WK = Word Knowledge; AR = Arithmetic Reasoning; DI = Data Interpretation; MK = Math Knowledge; RB = Rotated Blocks; EM = Electrical Maze; HF = Hidden Figures; TR = Table Reading; SR = Scale Reading; IR = Instrument Comprehension; AI = Aviation Information.

Figure 9. Bifactor SEM Model for (a) Sample 1 and 2 data; (b) Sample 3 data

Table 14 presents the standardized path coefficients of the predictive relations that resulted from separate analysis of nine different performance criteria. All tested SEM models fit the data well, as shown in Appendix B of the extended results. Clearly, across the performance criteria, only aviation-related acquired knowledge, general ability, and perceptual speed showed unique contributions in the predictive models.

Table 13. *Prediction of Pilot Performance by General Ability and Specific Abilities Across Three Pilots' Samples via Bifactor Models (Cross-validation Study)*

	Sample 1			Sample 2			Sample 3		
	Primary	Advanced	Rank	Primary	Advanced	Overall	Primary	Advanced	Academic
Verbal	$-.12^{+}$	$-.02^{ns}$	$-.05^{ns}$	$-.09^{***}$	$-.08^{+}$	$-.06^{*}$	$-.21^{***}$	$-.30^{***}$	$-.11^{**}$
Quantitative	$.07^{ns}$	$.15^{ns}$	$.22^{ns}$	$.07^{ns}$	$-.18^{ns}$	$.00^{ns}$	$-.03^{ns}$	-.07+	$-.01^{ns}$
Spatial	$-.03^{ns}$	$.10^{ns}$	$.07^{ns}$	-.07+	$-.0^{ns}$	$-.02^{ns}$	-	-	-
Perceptual	$.13^{ns}$	$.20^{*}$	$.26^{*}$	$.12^{**}$	$.10^{ns}$	$.17^{**}$	-	-	-
Knowledge	$.19^{***}$	$.23^{***}$	$.28^{***}$	$.34^{***}$	$.15^{***}$	$.30^{***}$	$.11^{***}$	$.09^{*}$	$-.01^{ns}$
g Bifactor	$.12^{ns}$	$.04^{ns}$	$.12^{ns}$	$.21^{***}$	$.14^{*}$	$.18^{***}$	$.26^{***}$	$.30^{***}$	$.52^{***}$

Note. gray color indicates negative estimate or nonsignificant positive estimate ($p < .05$). Models fit were all acceptable.
ns > .10. + < .10. * $p < .05$. ** $p < .01$. *** $p < .001$.

In Sample 1, aviation-related acquired knowledge was the only significant predictor of primary flight performance (β = .19, $p < .001$), and was coupled with perceptual speed as the only significant predictors of advanced flight performance (Knowledge: β = .23, $p < .001$; Perceptual: β = .20, $p < .05$) and students' class rank (Knowledge: β = .28, $p < .001$; Perceptual: β = .26, $p < .05$). Surprisingly, the general factor, as well as quantitative and spatial abilities did not show any noteworthy effect on any of the three criteria when controlling simultaneously for all these abilities. In Sample 2, aviation-related acquired knowledge remained the best predictor of the three criteria (β = .34; .15; .30, $p < .001$, for performance in primary, advanced, and overall training, respectively), followed by general ability (β = .21; .14; and .18, $p < .05$, for primary, advanced, and overall training, respectively). Perceptual speed had two significant effects, on primary flight performance (β = .12, $p < .01$) and students' rank (β = 17, $p < .01$) criteria. In Sample 3, general ability significantly predicted three performance measures (β = .26, .30, and .52, $p < .001$, for primary, advanced, and academic, respectively), and aviation-related acquired knowledge significantly predicted two measures (β = .11 and .09, $p < .05$ for primary and advanced), while verbal and quantitative abilities did not relate uniquely with any of the criteria.

4.3.5 Summary

This study contributes to the ongoing discussion about the relative importance of general and specific abilities for predicting job performance criteria. It further validates the result obtained in the primary study of this thesis across multiple criteria. A unique feature in this study is that it examined the validity of ability factors for predicting broadly versus narrowly defined criteria of performance, using a SEM approach that is appropriate for such analysis. Through analyzing three more pilots' samples, many results revealed in the primary study were replicated in this cross-validation effort, despite a few unexpected findings. Results generally supported the significance of general ability as a valid predictor for some general (e.g., overall composites of training outcome, overall academic achievement) and specific (e.g., flying performance in primary and advanced phases of training) performance criteria, with a few nonsignificant relations (e.g., the three measures in Sample 1). After separating domain-general from domain-specific abilities, the same two abilities emerged as important predictors of pilot performance in the primary study (i.e., acquired aviation knowledge and perceptual speed), also appeared the most contributing ability to the prediction of performance criteria in this study, with unique variance. Their role appeared even greater than that of *g* for several outcomes. Also, it is notable that the predictive relation pattern between specific abilities and pilot performance differed from their correlational pattern, as many of the significant relations in the correlational models turn out to be insignificant in the predictive models, suggesting that *g* factor fully accounted for these apparent effects. This was particularly true for the constructs of quantitative and spatial abilities, which gives an indication of the influence of these two constructs on psychometric *g*.

Overall, in the presence of general ability, the interplay of cognitive abilities in predicting pilot performance narrowed down to only three abilities out of seven potential ability predictors. This suggests that general ability absorbed a large portion of the common variance attributed to specific abilities and hence, emerged as a significant predictor. However, aviation-related acquired knowledge, in particular, related uniquely with all criteria across four studies (validation and cross-validation), suggesting its effect is even larger than that of general ability in predicting flight performance, at least for the actual flying aspect of performance, which is arguably the most important desirable outcome in training programs.

4.4 Cross-occupation Validation Study

The objective of this study was to evaluate the consistency of the predictive model across aviation-related occupations. Cognitive ability requirements often differ from one occupation to another. Thus, the comparison between jobs in ability-performance relationships can be beneficial to understand how these jobs are different and which tasks are most useful for selection of trainees in each. Specifically, the role of cognitive abilities in predicting training performance in three aviation jobs: flying, navigation, and air battle management (ABM), were examined in this study. The same four-step procedure followed in the previous two studies was also employed here, including deciding how to model performance measures, testing CFA measurement models, assessing the correlations between ability factors and performance measures, and applying bifactor SEM models to each data set.

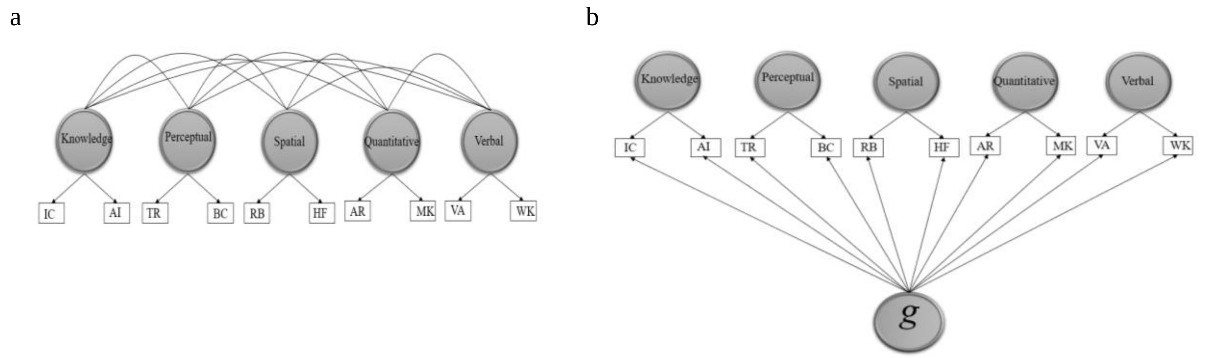
4.4.1 Modeling Performance

Data of pilots and navigators were obtained from Olea and Ree’s (1994) study, while ABM data was obtained from Carretta’s (1997) study, which all included intercorrelation matrices for AFOQT subtest scores and flight performance measures. Pilots’ performance was indicated by three measures: Pass/Fail training outcome, primary phase overall composite score, and advanced phase overall composite score. Navigators’ performance was indicated by three measures:

Pass/Fail training outcome, day check flight, and night check flight. For ABM, performance measure was an average final score on several written tests taken during the training program. Pilot and navigator performances were modeled as latent factors, whereas ABM performance retained as an observed measure.

4.4.2 Modeling Cognitive Abilities

Measurement models of cognitive abilities represented by five factors were fit to each of the three datasets. Due to the use of AFOQT Form S (11 subtests) in the ABM sample, it was only possible to manifest each ability construct with two indicators. In order to allow a direct comparison across the three samples, a modified version for the five-factor model was specified and tested. Figures 8a and 8b portray the two models used in this study. Each construct was indicated by: VA and WK for verbal, AR and MK for quantitative, HF and RB for spatial, and TR and BC for perceptual ability, and IC and AI for aviation-related acquired knowledge. With the exception of the perceptual speed factor, all other factors were manifested by subtests that appeared in some previous analyses. Because the Scale Reading subtest was omitted from AFOQT Form S, I coupled the Table Reading (TR) subtest with the Block Counting (BC) subtest to construct a perceptual speed factor. BC is commonly suggested as having a primary loading on spatial factors and a secondary loading on perceptual speed factors, which was also supported by the meta-analytic result of this thesis (preliminary study). Hence, construct validity of the perceptual speed factor can be adequately achieved by a combination of scores from these two tests (TR and BC).



Note. VA = Verbal Analogies; WK = Word Knowledge; AR = Arithmetic Reasoning; MK = Math Knowledge; RB = Rotated Blocks; HF = Hidden Figures; TR = Table Reading; BC = Block Counting; IR = Instrument Comprehension; AI = Aviation Information.

Figure 10. Cross-occupation study (a) correlated-factor model; (b) Bifactor model.

The CFA measurement models for each sample (correlated-factor and bifactor) and the models' fit statistics are reported in Appendix B. All models fit the data adequately. As seen in Table 15, all factor loadings from the correlated-factor models were significant across the three samples ($p \leq .001$). For each sample, the Table Reading subtest on perceptual speed factor had the weakest loading (.48–.51), whereas the Verbal Analogies subtest on the verbal ability factor had the largest loading (.91–.94). The weakest factor correlations across the three samples were those between aviation acquired knowledge and quantitative ability (.19–27), while the largest factor correlations were those between spatial ability and perceptual speed (.67–.76).

Table 14. *Factor Loadings from Correlated-factor and Bifactor Models (Cross-occupation Validation Study)*

Sample	Model	Factor	Verbal		Quantitative		Spatial		Perceptual		Knowledge	
			VA	WK	AR	MK	RB	HF	TR	BC	IC	AI
Pilots	Correlated-factor		.94	.60	.82	.69	.51	.56	.48	.75	.62	.64
	Bifactor	Specific	.80	.50	.51	.40	.27	.30	.31	.49	.59	.65
		General	.49	.38	.61	.62	.51	.41	.40	.55	.38	.05
Navigators	Correlated-factor		.91	.60	.86	.65	.51	.54	.48	.79	.67	.64
	Bifactor	Specific	.78	.50	.65	.48	.20	.23	.31	.51	.58	.59
		General	.50	.37	.55	.47	.54	.43	.38	.59	.46	.17
Air Battle Managers	Correlated-factor		.91	.70	.81	.76	.69	.65	.51	.87	.67	.69
	Bifactor	Specific	.78	.59	.59	.60	.27	.26	.34	.61	.52	.59
		General	.48	.39	.58	.43	.63	.60	.43	.60	.56	.29

Note. All loadings were significant at $p < .001$, except for the Aviation Information subtest on *g* factor in the pilot sample ($p = .07$).

For the bifactor models, the standardized loading coefficients of the ten subtests on both *g* factor and specific ability factors are presented in Table 15. Even in the presence of *g*, the five ability factors in the three samples remained clearly evident with significant loadings. Across samples, the factor that seemed most greatly influenced by the presence of *g* in the models was the spatial ability factor, as indicated by the weak loadings of its two indicators (.20 to .30). The lowest loadings on the general factor were those produced by the Aviation Information subtest, one of the two indicators of the aviation-related acquired knowledge factor (.05, $p = .07$ for pilots' sample; .17, $p < .001$ for navigators' sample; .29, $p < .001$ for ABM sample). Notably, the Arithmetic Reasoning subtest did not seem to be as strongly related to the *g* factor as in the former analyses. Overall, this initial phase confirmed the soundness of the suggested five-factor model for aviation trainee cognitive abilities.

4.4.3 Correlation-based Validity

Correlated-factor models associating the five cognitive abilities with job performance criteria were then specified. The CFA fit statistics of this model for the three samples were all acceptable, as presented in Table 16. All factors correlated substantially with the latent variable of flying performance, with the exception of the verbal ability factor (r = -.01, p > .10). The magnitudes of the significant relations ranged from .11 (quantitative ability) to .32 (aviation acquired knowledge). For navigators, the relations of cognitive abilities with latent performance outcomes were all significant with no exception, ranging from .13 to (verbal ability) to .40 (spatial ability). For ABMs, the correlations between the five abilities and the observed performance variables were also all significant, ranging between .19 (perceptual speed) and .32 (quantitative ability). This phase of analyses indicated that associations between cognitive abilities and training performance were generally significant, although the pattern differed across occupations.

Table 15. *Correlations Between Cognitive Abilities and Job Performance (Cross-occupation Validation Study)*

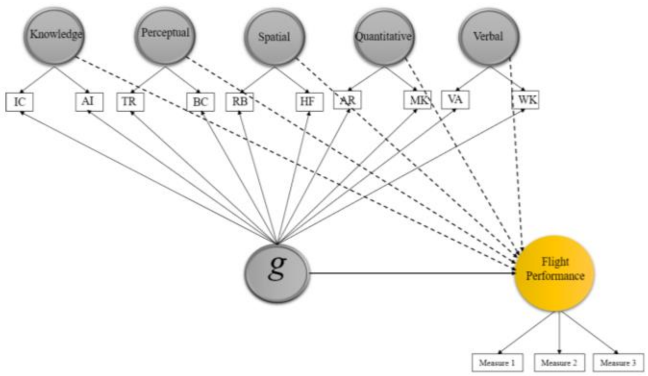
	Flying	Navigation	Air Battle Management
Verbal Ability	$-.01^{ns}$	$.13^{**}$	$.29^{***}$
Quantitative Ability	$.11^{***}$	$.37^{***}$	$.32^{***}$
Spatial Ability	$.15^{***}$	$.40^{***}$	$.22^{***}$
Perceptual Speed	$.17^{***}$	$.32^{***}$	$.19^{***}$
Aviation Acquired Knowledge	$.32^{***}$	$.16^{***}$	$.29^{***}$

Note. gray color indicates negative estimate or nonsignificant positive estimate (p > .05). Models fit were as follows. Flying: $\chi^2(55)$ = 558.47, p < .001; CFI = .93; RMSEA = .07; SRMR = .06; Navigation: $\chi^2(55)$ = 186.63, p < .001; CFI = .94; RMSEA = .05; SRMR = .05; ABM: $\chi^2(35)$ = 158.68, p < .001; CFI = .94; RMSEA = .07; SRMR = .06.
ns p > .10. ** p < .01. *** p < .001.

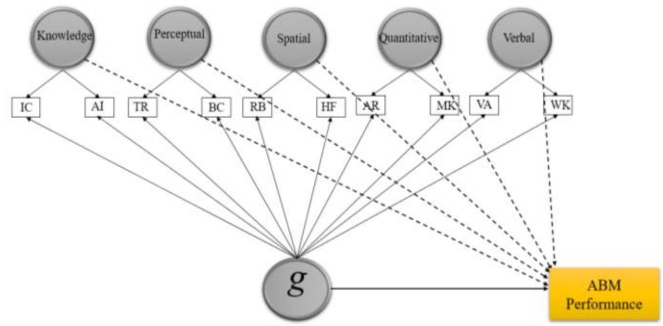
4.4.4 SEM-based Validity (SEM Predictive Model)

In this phase, a bifactor modeling approach was used to investigate the predictive value of *g* and domain-specific abilities. Figures 9a displays the structural model used for the pilot and navigator samples. Figure 9b displays the same for the ABM sample. Models fit and path coefficients resulting from the predictive bifactor models are shown in Table 17. The models fit the three correlation matrices well (pilots: CFI = .94; RMSEA = .07; SRMR = .05; navigators: CFI =.95; RMSEA =.05; SRMR = .04; battle managers: CFI =.95; RMSEA =.07; SRMR = .05).

a



b



Note. Solid lines represent paths from the general ability factor; dashed lines represent paths from the specific ability factors. VA = Verbal Analogies; WK = Word Knowledge; AR = Arithmetic Reasoning; MK = Math Knowledge; RB = Rotated Blocks; HF = Hidden Figures; TR = Table Reading; BC = Block Counting; IR = Instrument Comprehension; AI = Aviation Information.

Figure 11. Bifactor SEM Model for (a) Pilots and Navigators data; (b) ABM Sample

Table 16. *Prediction of Job Performance by General Ability and Specific Abilities Via Bifactor Models (Cross-occupation validation Study)*

	Flying	Navigation	Air Battle Management
Verbal Ability	$-.07^{ns}$	$-.14^{ns}$	$.24^{**}$
Quantitative Ability	$.07^{ns}$	$.15^{ns}$	$.32^{**}$
Spatial Ability	$.04^{ns}$	$.02^{ns}$	$.33^{ns}$
Perceptual Speed	$.10^{ns}$	$-.01^{ns}$	$.18^{ns}$
Aviation Acquired Knowledge	$.29^{***}$	$-.05^{ns}$	$.31^{**}$
g (Bifactor model)	$.11^{*}$	$.42^{**}$	$.10^{ns}$

Note. gray color indicates negative estimate or nonsignificant positive estimate ($p > .05$). Models fit were as follows. Flying: $\chi^2(54) = 494.52$, $p < .001$; CFI = .94; RMSEA = .07; SRMR = .05; Navigation: $\chi^2(54) = 162.03$, $p < .001$; CFI = .95; RMSEA = .05; SRMR = .04; ABM: $\chi^2(34) = 141.22$, $p < .001$; CFI = .95; RMSEA = .07; SRMR = .05
ns $p > .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Results showed interesting findings concerning the predictive relation of general ability and job performance. Among the three samples, the *g* effect was notably high only in the navigation sample, where its estimate was the only significant estimate among predictors in the model, with relatively high beta ($\beta = .42$, $p < .01$). Conversely, the *g* effect was small and not significant in the ABM sample ($\beta = .10$; $p > .05$), whereas three specific abilities emerged as strong predictors: quantitative ability, aviation acquired knowledge, and verbal ability ($\beta = .32$, .31, .24, respectively; $p < .01$). For the pilot sample, the *g* effect was rather small and barely reached the significance level ($\beta = .11$; $p < .05$), whilst aviation acquired knowledge, the only meaningful specific predictor in the model, predicted flight performance much better ($\beta = .29$, $p < .001$).

4.4.5 Summary

This study attempted to provide further evidence from aviation for the predictive relations between cognitive abilities and job performance. Using the same bifactor predictive model (five abilities), results clarified the interplay of general and specific cognitive abilities in predicting performance in three aviation jobs: flying, navigation, and air battle management (ABM). First, the effect size of bifactor *g* was large in the navigation sample, small in the flying sample, and negligible in the ABM sample. In contrast, the number of significant effects due to specific factors was none in the navigation sample, one in the flying sample, and three in the ABM sample. In the navigator sample, the number of effects that were uniquely attributable to specific abilities in the bifactor model was almost trivial. When *g* was modeled, the effect of specific abilities either declined or faded away, as can be seen from the nonsignificant effects of most predictors (Table 17) as compared to their significant simple correlations with performance criteria (Table 16). This suggests that the simple correlations of the five abilities with navigation performance ($r = .13$ to $.40$) were mostly due to their overlap with *g*, not unique variance of these abilities.

However, the conclusion drawn for the navigators' sample might not apply to the other two samples. In fact, the ABM sample showed a contrasting interplay of cognitive abilities, where some specific abilities "resisted" the entry of *g* in the bifactor model such that their prediction role remained significant, even greater than *g*'s predictive role. This implies that domain-specific factors in this sample (i.e., quantitative, perceptual, and verbal abilities) had a unique effect above and over the general-domain ability, which gives evidence for their important contribution to performance in this job. The pattern in the pilots' sample was intermediate between these two contrasting patterns, where the aviation acquired knowledge, along with *g*, stayed significant and effective in the predictive model.

4.5 Combined Results from the Predictive Studies

In order to facilitate the discussion of the findings, I present here a general summary of the results concluded from the three predictive studies. The below summary tables include four tables: factor loadings resulted from the CFA correlated-factor and bifactor models, factors intercorrelations resulted from the CFA correlated-factor models, predictive validity estimates of cognitive abilities resulted from their combined correlated-factor models with job performances, and the effects of cognitive abilities on job performances resulted from bifactor SEM models. Because studies included four to five data points for pilot, Tables 17 and 18 show only the averages of the statistics, and the complete summary can be found in Appendix B.

Table 17. *Summary for Factor Loadings Based on CFA Correlated-factor and Bifactor Models*

Subtest	Correlated-factor	Bifactor	
		Specific Factors	General Factor
VA	**.78;** *.91;* .91	**.56;***.78;* .78	**.53;** *.50;* .48
RC	**.80**	**.58**	**.52**
WK	**.74;** *.60;* .70	**.60;** *.5;* .59	**.47;** *.37;* .39
AR	**.77;** *.86;* .81	**.32;** *.65;* .59	**.66;** *.55;* .58
DI	**.73**	**.24**	**.72**
MK	**.67;** *.65;* .76	**.43;** *.48;* .6	**.60;** *.47;* .43
RB	**.57;** *.51;* .69	**.43;** *.20;* .27	**.42;** *.54;* .63
EM	**.49**	**.37**	**.32**
HF	**.49;** *.54;* .65	**.32;** *.23;* .26	**.34;** *.43;* .60
TR	**.55;** *.48;* .51	**.36;** *.31;* .34	**.38;** *.38;* .43
SR/BC	**.69;** *.79*;* .87*	**.42;** *.51*;* .61*	**.59;** *.59**; .60*
IC	**.64;** *.67;* .67	**.59;** *.58;* .52	**.36;** *.46;* .56
AI	**.63;** *.64;* .69	**.60;** *.59;* .59	**.15;** *.17;* .29

Note. bold font represents average estimates of four to five correlations from pilot samples; italicized font represents estimates from navigator sample; normal font represents estimates from air battle manager sample.

Table 18. *Summary for Factor Intercorrelations Based on CFA Correlated-factor Models*

	Verbal	Quantitative	Spatial	Perceptual	Knowledge
Verbal	1				
Quantitative	**.64**; *.50*; .52	1			
Spatial	**.38**; *.46*; .46	**.55**; *.54*; .55	1		
Perceptual	**.34**; *.34*; .31	**.71**; *.47*; .50	**.67**; *.76*; .67	1	
Knowledge	**.29**; *.29*; .32	**.31**; *.26*; .27	**.46**; *.55*; .67	**.32**; *.36*; .42	1

Note. bold font represents average estimates of four to five correlations from pilot samples; italicized font represents estimates from navigator sample; normal font represents estimates from air battle manager sample.

For the sake of comparison between validity estimates for specific outcome criteria with which cognitive abilities were associated, Tables 19 and 20 were arranged according to four distinct criteria: hands-on flying performance at the primary phase of training (four data points), hands-on flying performance at the advanced phase of training (four data points), overall index of performance (three data points), and academic performance throughout training program (one to two data points). Both tables also include average scores representing these cognitive ability estimates across criteria.

Table 19. *Combined Results for the Correlations Between Cognitive Abilities and Performance Measures Grouped by Four Performance Criteria*

										Navigator	**ABM**
	Primary	***M***	Advanced	***M***	Overall	***M***	Academic	***M***	***M***		
Verbal	.06, -.02, .04, .07	**.04**	-.02, .01, .00, .05	**.01**	.04, .05, -.01	**.03**	.19, .35	**.27**	**.09**	.14	.31
Quantitative	.20, .13, .19, .21	**.18**	.10, .10, .09, .21	**.13**	.20, .16, .10	**.15**	.26, .46	**.36**	**.21**	.39	.30
Spatial	.15, .08, .14	**.12**	.06, .12, .12	**.10**	.15, .17, .15	**.16**	.12	**.12**	**.13**	.40	.22
Perceptual	.37, .18, .26	**.27**	.20, .17, .19	**.19**	.27, .16, .18	**.20**	.23	**.23**	**.22**	.33	.22
Knowledge	.49, .22, .40, .21	**.33**	.16, .22, .20, .21	**.20**	.30, .26, .19	**.25**	.13, .26	**.20**	**.25**	.14	.21
M		**.19**		**.13**		**.16**		**.24**			

Note. Primary = performance measures collected during the primary phase of training; Advanced = performance measures collected during the advanced phase of training; Overall = overall index measuring several dimensions of performance; Academic = performance of trainee in flying courses taught in ground school; *M* = Mean; ABM = Air Battle Manager; bold color = the mean of the estimates.

Table 20. *Combined Results for the Effects of Cognitive Abilities on Performance Measures Grouped by Four Performance Criteria*

	Pilot									**Navigator**	**ABM**
	Primary	***M***	Advanced	***M***	Overall	***M***	Academic	***M***	***M***		
Verbal	-.14, -.12, -.9, -.21	**-.14**	-.10, -.02, -.08, -.30	**-.13**	-.05, -.06, -.07	**-.06**	.04, -.11	**-.04**	**-.09**	-.14	.24
Quantitative	-.07, .07, .07, -.03	**.01**	-.02, .15, -.18, -.07	**-.03**	.22, .00, .07	**.10**	.08, -.01	**.04**	**.03**	.15	.32
Spatial	-.09, -.03, -.07	**-.06**	-.03, .10, .00	**.02**	.07, -.02, .04	**.03**	-.06	**-.06**	**-.02**	.02	.33
Perceptual	.23, .13, .12	**.16**	.16, .20, .10	**.15**	.26, .17, .10	**.18**	.06	**.06**	**.14**	-.01	.18
Knowledge	.43, .19, .34, .11	**.27**	.12, .23, .15, .09	**.15**	.28, .30, .29	**.29**	.08, -.01	**.04**	**.19**	-.05	.31
g bifactor	.26, .12, .21, .26	**.21**	.12, .04, .14, .30	**.15**	.12, .18, .11	**.14**	.24, .52	**.38**	**.22**	.42	.10

Note. Primary = performance measures collected during the primary phase of training; Advanced = performance measures collected during the advanced phase of training; Overall = overall index measuring several dimensions of performance; Academic = performance of trainee in flying courses taught in ground school; *M* = Mean; ABM = Air Battle Manager; bold color = the mean of the estimates.

4.6 General Summary

This thesis offered four important investigations, one preliminary study to assess the test battery that was used for sourcing cognitive data and three predictive studies to assess the role of cognitive abilities as predictors for pilot (and two other aviation jobs) performance. The assessment of five datasets comprising 13 total pilot performance outcomes, observed and latent, clearly indicated that aviation-related acquired knowledge is the most important construct in the prediction of future flight performance. This factor shows robustness in its effects on performance criteria, even when the effect of general ability is also accounted for. More importantly, the effect size of aviation-related acquired knowledge exceeded the effect of general ability in several cases, suggesting that this construct, as indexed by the subtests of Instrument Comprehension and Aviation Information, is a valuable construct for future pilot training performance and warrants inclusion in selection test batteries that are designed for pilot recruitment. The perceptual speed construct is another noteworthy predictor of pilot performance, with several unique contributions to prediction, even when general ability present in the models. Three of ten examined effects of this construct were found not significant, suggesting being thoughtful about the performance measures to which this construct is associated.

General cognitive ability plays an important role in flying performance similar to that revealed in many other jobs. Based on the bifactor modeling of several datasets, quantitative ability and spatial ability tend to be absorbed in the *g* factor. Such a pattern is evident when the bifactor model is extended to a predictive SEM model, where both abilities failed to show any incremental validity beyond that produced by the *g* factor, although both showed significant relationships with performance measures in the correlational models. This pattern is not so clear for verbal ability because its significant correlations with performance measures, in the first place, were few in number and its unique predictive role was established for only one of 13 criteria (i.e., academic performance). Moreover, the significant negative effects noted in several instances for verbal ability with pilot performance may suggest that a higher verbal aptitude can be even undesirable for flying outcomes. Because using a mix of verbal, quantitative, or spatial abilities is a common strategy in estimating psychometric *g* (e.g., Ree et al., 1995), this result has an important implication for pilots' selection. Given the trivial unique roles found for verbal, quantitative, and spatial abilities as predictors for pilot performance, a score of *g* obtained from a single reliable measure of general ability (e.g., Raven's Progressive Matrices) may suffice to predict trainee outcomes, and multiple tests may not be necessary.

The comparison between pilots and the other two aviation jobs (navigators and air battle managers) showed a different interplay between cognitive abilities in predicting job performance. General ability effects were large on navigators' performance, small on pilots' performance, and unimportant in ABM performance. Conversely, the number of specific effects that uniquely predicted job performance (after accounting for the effect of *g*) were none for navigators' performance, one for pilots' performance, and three for ABM performance. These findings support the notion that each job requires a unique set of cognitive abilities, which may be different from those needed for other occupations. Aviation-related acquired job knowledge continued to show a focal role in the prediction of subsequent pilot training performance. A limitation of this comparison may be the type of ABM performance measure used in the study, which represented an academic performance, not hands-on ABM performance. The comparison would have been more reliable if the data had included work sample measures, in addition to the academic performance measure.

CHAPTER 5: GENERAL DISCUSSION

Intelligence researchers have long debated whether the general ability factor is the only factor that accounts for the performance in cognitive tasks or there might be other broad ability factors that explain some of the common variance in test scores (e.g., Agnello et al., 2015; Reeve & Bonaccio, 2011). Another version of this debate is the debate among industrial/organizational (I/O) psychology researchers about whether it is general ability or narrower abilities that contribute most to the prediction of job and training performance (e.g., Hunter, 1986; Kell & Lang, 2017; Lang et al., 2010; Ones, Viswesvaran, & Dilchert, 2012). The current thesis weighs in on this controversy by providing results that may be of mutual interest to intelligence and I/O psychology researchers using data from highly cognitively demanding occupations, where individual differences in job performance are linked to differences in cognitive abilities.

The predictive relations between cognitive abilities and flight performance measures were thoroughly investigated in this thesis using four pilot samples who were tested on the Air Force Officer Qualifying Test (AFOQT) during the selection process and rated on several specific performances during flight training. Because it is important to align the level of a hypothesis (observed versus latent variables) with the level of analysis (observed versus latent variables) (Ullman & Bentler, 2012), this study used a SEM approach to investigate the relative importance of general and specific abilities in predicting pilot flying performance. Two types of predictive models were estimated in the thesis: correlation-based and bifactor SEM-based. Correlation-based models correspond to the common practice in validation research that relies on simple correlations between ability and performance scores, whereas bifactor SEM-based models may better correspond to the explanatory research that seeks a theoretical understanding of the relations (e.g., Pedhazur, 1997). In order to compare the role of cognitive abilities in pilots' performance with their role in other aviation-related jobs, data from navigators and air battle managers were also modeled. This should provide further understanding of ability-performance relationships across aviation occupations.

The discussion in this chapter is outlined to highlight the following points: (1) The AFOQT factor structure (EFA models, CFA correlated-factor models, and CFA bifactor models), (2) correlation-based predictive validity, (3) the effects of cognitive abilities on pilot performance, (4) cognitive abilities across three aviation occupations, (5) the added value of the current results, (6) implications and recommendations, and (7) limitations and future research, (8) conclusion.

5.1 The AFOQT factor structure

The four studies reported in this thesis provide a useful assessment for the AFOQT factor structures by means of both EFA and CFA approaches. In addition to the exploratory approach employed in the preliminary study to assess the AFOQT data meta-analytically (i.e., EFA and second-order EFA), CFA correlated-factor and CFA bifactor models were applied before testing SEM predictive models of pilot performance. Representation of the AFOQT data using these two models can provide evidence for the most plausible factor structure of its subtest data. It can also be informative for the recently emerging line in psychometric research to compare between several competing models in the attempt to understand the best representation of human cognitive structure (e.g., Cucina & Byle, 2017; Morgan et al., 2015; Murray & Johnson, 2013).

5.1.1 EFA Model

The six factor structures assessed in the preliminary study via EFA yielded a useful understanding of the ability factors underlying AFOQT. What makes this examination of particular interest is that the EFA was applied to pooled correlation matrices of pilot cognitive performance data, spanning decades of AFOQT research. The evidence resulting from this study thus may be more reliable and stronger than that derived from any single set of data. I conclude that the five- and four-factor models are more plausible structures for the AFOQT data than the one-, two-, three-, and six-factor structures. Both yielded acceptable fit indices and had readily interpretable, distinct ability factors. However, the joint evaluation of model selection criteria (e.g., model fit statistic) in this thesis provides more support for the five-factor model that also resemble the AFOQT's theorized factor structure. Due to the existence of only one EFA investigation for the 16 AFOQT subtests (i.e., Skinner & Ree, 1987), in addition to an EFA for 12 subtests (Johnson et al., 2017), the present results examining multiple AFOQT factor structures can be useful.

The five-factor AFOQT structure that emerged in the current study is almost identical to the five-factor structure that was suggested by Skinner and Ree (1987). They suggest that AFOQT is best described by five ability factors tapping verbal, quantitative, and spatial ability, perceptual speed, and aviation-related acquired knowledge. One notable and one minor difference between the current and prior results can be highlighted. The notable difference is related to the lack of significant loading of the General Science subtest on the aviation acquired knowledge factor in this study, while it was suggested as a primary indicator for this factor in the previous EFA (Skinner & Ree, 1987). The minor difference is related to the stronger loading of the Scale Reading subtest on the quantitative ability factor relative to its weaker loading on the perceptual speed factor in this study (although both are significant: .42 and .35, respectively), whilst the former EFA suggested that it is a primary indicator for perceptual speed and a secondary indicator for quantitative ability. Performance in this particular subtest involves several cognitive functions (e.g., visual processing, processing speed, and fluid-quantitative reasoning; see Table 2), which may explain this variation in its factor loading. The construct underlying this subtest may even approximate the construct of general ability, which is typically captured through such a mix of cognitive function. These few differences between the current EFA and previous EFA may also be attributed to the characteristics of the current sample (i.e., pilot sample) as compared to the general officer applicants sample used in Skinner and Ree's (1987) study. The four-factor structure suggested as a plausible alternate population model in this study shows a similar latent ability factor to that of the five-factor structure, but with one common factor comprising the subtests of quantitative ability and perceptual speed factors.

Johnson et al. (2017) suggested a two-factor solution for 12 of the 16 AFOQT subtests (after excluding IC, AI, GS, and DI) from a sample of 30,025 USAF officer applicants. This was a result of principal components analysis with Varimax rotation, and informed by two criteria for the number of factors to be retained (scree plot and factors with eigenvalues > 1 criteria). Although the preliminary study of this thesis applied a different EFA procedure (i.e., principal axis factoring with promax rotation) for a subpopulation from the general officer applicants population (i.e., pilot) and determined a different conclusion about the plausibility of the two-factor model (i.e., unlikely), results yielded a pattern of factor-subtest relations for the two-factor solution that is fairly comparable to Johnson et al. (2017). Specifically, the suggested interpretation of the two factors as Spatial Cognition and Academic Aptitude seems also likely for the two factors that emerged in the current EFA, even with the inclusion of the four excluded subtests (IC, AI, GS, DI). The primary content of Spatial Cognition factor in Johnson et al. (2017) was MC, EM, BC, RB, HF, SR, and TR, and the same set of subtests, in addition to the SR and IC, are the primary content of the corresponding factor in this study. Similarly, the primary content of Academic Aptitude factor in Johnson et al. (2017) was VA, RC, and WK, and the same set of subtests, in addition to the GS, are the primary content of the corresponding factor in this study. Two of three subtests cross-loaded on both factors in Johnson et al. (2017) (AR, MK, and MC) are also cross-loaded in the current EFA (AR, MK, and DI). The AI subtest in this study (excluded in the previous study) did not load significantly on either of the two factors, likely resulting from its domain-specific content that is related to flying knowledge.

The assessment of factors' contents and factors' intercorrelations in the tested solutions revealed some information about constructs' relations at the broad level of abilities. For example, the five-factor models shows strong relations for quantitative ability with verbal and spatial abilities, and weak relations for perceptual speed with verbal ability and aviation-related acquired knowledge. The four-factor structure suggested that the two constructs of quantitative ability and perceptual speed ability have considerable common variance to share. The two-factor solution showed a divergent relationship between verbal ability and spatial-perceptual cognition factor and a convergent relationship between quantitative ability and both spatial-perceptual cognition and academic factors. The single-factor model indicates that quantitative ability subtests and, to some extent, verbal ability subtests are important elements in the construction of psychometric *g* ($\lambda \geq .60$), while domain-specific acquired knowledge seem less relevant to *g*. This pattern of relations between specific constructs and the general ability was further confirmed when attempting second-order EFA for the five-factor model that exhibited stronger loadings of quantitative and verbal ability subtests on the *g* factor, as compared to the lower loadings of aviation-related acquired knowledge subtests.

5.1.2 CFA Correlated-factor Model

A CFA correlated-factor model was fit to each data set examined in this thesis. When constraining the general factor loadings of the bifactor model to zero and relaxing the orthogonality constraint on the first-order factors, the correlated factors model can thus be derived. The correlated-factor model is one popular model for hypothesizing the structure of cognitive data (e.g., Carroll, 1997). Studies consistently show that these models possess strengths and tended to fit the cognitive data adequately. For example, Morgan et al. (2015) found in their simulation study of three competing solutions that correlated factors solution is the best fitting model when samples are selected from a true multiple correlated factors structure, based on approximate fit indices.

Across seven separate analyses, the CFA correlated-factor model with five group factors was found to be an acceptable representation for the AFOQT data. This conclusion seems similar to those of Carretta et al.'s (2016) recent study, where a five-correlated factor model, along with a higher-order factor model with five content factors, were found to be more likely as compared to single-factor, four-correlated factor, and higher-order four factor models. However, Carretta et al.'s (2016) analyses were performed using item-level data and the item-parceling technique, so the results are not directly comparable to those of the present study. The pattern of relations between ability factors is worth highlighting. The strongest association was noted between quantitative ability and perceptual speed. This notable relationship between the two constructs in AFOQT may be due to the influence of the SR subtests that have common content shared by the two constructs. It was also demonstrated that spatial ability-perceptual speed, quantitative ability-verbal ability, and quantitative ability-spatial ability are related strongly to each other. Verbal ability relation with spatial ability was weaker than its relations with the quantitative ability or perceptual speed. Not surprisingly, domain-specific acquired knowledge related somewhat weakly with some of the other ability factors; particularly, with verbal ability and quantitative ability. Although this pattern of relations is offered in the context of AFOQT data, a similar pattern may be expected in the true relationships between constructs (e.g., Lang et al., 2010; Legree et al., 1996; Paunonen & Hong, 2010; Wee et al., 2014).

5.1.3 CFA Bifactor Model

Research focusing on either the unique aspects of specific abilities (e.g., verbal, quantitative) or the common factor (e.g., *g*) can be limited and may not convey a complete picture of human cognition. In contrast, the bifactor conceptualization of ability structure is useful as it reconciles the varying perceptions and incorporates in its framework both common and unique aspects of cognitive abilities (McFarland, 2016). This thesis adds seven analyses of AFOQT data to the growing body of evidence supporting a bifactor structure of cognitive test batteries (Morgan et al., 2015; Murray & Johnson, 2013). The findings suggest that the structural relationship between the five AFOQT cognitive abilities can plausibly be represented by a bifactor model. Such conceptualization of *g* implies that it is more of a breadth factor, rather than a superordinate factor (Gignac, 2008). Because the extraction of *g* and the first-order factors from the observed subtest indicators are undertaken simultaneously, and at the same abstraction level, the models, therefore, are more parsimonious and less complicated than higher-factor models (Canivez, 2016; Cucina & Byle, 2017; Gignac, 2008).

Fitting several bifactor models with five group factors in this thesis indicated acceptable fit statistics and thus suggesting a viable factor structure of the AFOQT data. This result is consistent with prior AFOQT factor-analytic investigations that found bifactor models are possible representations of AFOQT data. Carretta and Ree (1996) found a bifactor model containing the five group factors suggested by Skinner and Ree (1987) as the best-fitting model among seven examined models. Using item-parceling, Drasgow et al. (2010) also showed that the AFOQT data are best described by a bifactor model containing the same five content-specific factors. Thus, the accumulated evidence tends to support preferences for the bifactor model over other competing models.

Some noteworthy findings have emerged from the bifactor models in this thesis that can be of theoretical interest, particularly pertaining to the subtest loadings on the general factor and the specific factors to which they belong. Across bifactor models, the AFOQT's quantitative ability subtests consistently loaded more firmly on the general factor than on their theorized specific factor. A reversed pattern was noted for the aviation-related acquired knowledge factor subtests (IC and AI), where their loadings on their corresponding specific factor were more substantial than their loadings on the general factor. The subtests of verbal ability, spatial ability, and perceptual speed produced a fairly balanced pattern of loadings on the general factor and their theorized specific factors. This pattern of subtest loadings on the general factor looks somewhat similar to that found in Earles and Ree's (1991) study, using different estimation methods (i.e., unrotated principal components, unrotated principal factors, and variants of hierarchical factor analysis). Consistent with the EFA result, quantitative ability seems to be a core component in the AFOQT's general ability formation, whereas domain-specific acquired knowledge is mostly orthogonal to general ability and contributes only modestly to its construct. In addition to three quantitative ability subtests (AR, DI, MK), the subtests of Scale Reading, Verbal Analogies, and Table Reading are influential and have a strong connection with *g*.

5.2 Correlation-based Predictive Validity

The three predictive studies in this thesis included an initial assessment for the associations between specific ability factors and performance measures (latent or observed). This early evaluation was useful because it demonstrated the extent to which ability scores measured in the selection process are related to pilot performance scores collected in different phases of the training program. It also showed the relative importance of specific abilities in predicting pilots' performance, before decomposing the variance to domain-general factor and domain-specific factors. Because correlation-based predictive validity remains a common procedure for assessing predictor-criterion relationships (e.g., Carretta, 2011; Carretta et al., 2014; King et al., 2013), this phase of analysis corresponds to a widely-used criterion-related validity investigation, but at construct level. Given that cognitive variables in these studies were modeled as latent factors, the findings determined might add value to those relying primarily on observed scores of ability tests. Interpretation at a broader construct level can thus be made more determinately (Brown, 2015).

Table 19 displays a summary of the results obtained in each study, as well as the averages. Two distinct patterns of associations between ability factors and pilot performance measures were identified; a pattern displayed by the associations with the first three criteria (flying and overall ratings) and a pattern shown by the associations with the academic performance of pilots. The first pattern of associations indicates that the relative importance of specific abilities tends to be consistent across the criteria of actual flying performance at both primary and advanced phases, as well as the overall performance criteria. Across these three criteria, aviation-related acquired knowledge was the factor most highly related to performance measures, with perceptual speed being the next most related factor with performances. Conversely, the verbal ability factor was least strongly related to the three performances. Quantitative and spatial abilities exchanged the third and fourth rank of relative importance. Nevertheless, inconsistent validity magnitudes were observed for the abilities across samples and criteria. For example, aviation-related acquired knowledge associations with the three performance criteria varied from .16 to .49 (Δr = .33), depending on the training phase with generally larger correlations with scores from primary phase measures than from advanced phase measures or overall performance. This variation might be expected due to the differences in measures collected in each study. For instance, the criterion from the primary training phase used in Sample 1 of the cross-validation study (Duke & Ree, 1996) was obtained by computing the differences between the flying hours needed by pilots to complete their primary training and the average of flying hours by all pilots in the sample. The criterion of the primary training phase used in Sample 2 of the cross-validation study (Olea & Ree, 1994) was an average rating of several check flights collected during this phase. Thus, the notable differences in the magnitude of ability factors across data sets could be a function of these notable differences in the scores representing performance criteria.

Another possibility for the variation in relations between predictor and criterion scores, although it is less clear, is the way that performance measures were modeled, either as an observed or latent variable. By way of illustration, performance in the primary phase of training was modeled as latent in two instances and as observed in two other instances. When data included an adequate number of measures in each performance dimension (e.g., academic, primary phase), the choice was to model these measures as indicating a latent performance factor. The quantitative ability factor, for example, was associated relatively more strongly with the latent performance of the primary phase than with observed performance. A similar observation was noted for verbal ability displaying a stronger relation with two latent performance factors than the corresponding observed variables. Due to the absence of measurement error in latent variables with which observed variables are contaminated, this result may be expected. For aviation acquired knowledge factor, however, no consistent pattern can be concluded in terms of its associations with latent/observed performances, because its validity estimates varied among the two types of modeling; .49 and .21 with latent factors of primary phase performance and .22 and .40 with the observed scores of performance.

Academic performance outcomes showed a different pattern of associations with cognitive abilities whereby quantitative and verbal abilities become remarkably influential predictors of academic performance, while spatial ability, perceptual speed, and acquired aviation knowledge become less important. Additionally, this is the only criterion for which verbal ability showed a significant relationship with pilot performance. This may suggest that although the verbal ability is not a good predictor for hands-on flying performance, it can still be a needed ability, with a positive influence on the academic aspect of training. Additionally, tests of verbal ability may also be required given its contribution in the general ability construct (e.g., Ree et al., 1995).

Further averaging of ability estimates across all criteria suggests that the associations with academic performance tend to generally exhibit larger validity estimates. This may be expected given the highly scholastic aptitude characterizing pilot selectees, which may be more relevant to performance in the academic aspect of the training program. The associations of abilities with performance scores at the primary phase of training are relatively higher than their associations with both the performance at the advanced phase of training and the overall performance. The strict scholastic and aptitude requirements in pilot selection procedures favor those possessing higher levels of general ability, as indexed by high scores on, for example, GPA, ACT, and AFOQT, which operate on specific prior knowledge (e.g., IC and AI subtests), and facilitate the acquisition of flying experience. This causal relationship between the general cognitive ability and prior job knowledge, which advance the prediction of job performance, may be more apparent at the initial phases of training, as evidenced by relatively higher correlation estimates between ability factors and performance criteria in this stage of training. The average associations of ability factors with a flying performance at the advanced phase appeared slightly weaker than the primary phase performance. Flying training, as in the case of most job training, progresses from teaching basic flying skills to more advanced skills, and hence, the attrition rate increases toward the end of the training program. In addition to cognitive abilities, many other factors contribute to the persistence of students in the training program and come into play at the advanced phases of training (e.g., motivation, personality trait) (Ackerman et al., 1995; Kanfer & Ackerman, 1989). Ackerman's (1988) model of skill acquisition may also explain this variation of validity for the same construct across performance criteria. According to Ackerman (1988), there are three main phases of skill acquisition, each of which requires different cognitive demands: the declarative phase requires more general ability, the knowledge compilation phase requires more perceptual speed ability, and the procedural phase requires more psychomotor ability. Hence, many of the AFOQT constructs may be more relevant to the first and second phases of Ackerman's (1988) model, while constructs representing psychomotor ability, which are not measured by AFOQT, may be more relevant to the type of performance suggested for an advanced phase of skill acquisition. Therefore, the expected relations between cognitive ability and advanced stages of training performance tend to be weaker relative to early phases.

Overall, the results from ability-performance correlations suggest a reasonable degree of consistency in terms of the relative importance of cognitive abilities as predictors for pilot performance (e.g., aviation acquired knowledge a better predictor than verbal ability). Less consistency, however, is suggested for the magnitude of validities across sample and performance criteria. Moreover, academic performance and hands-on flying performance criteria should always be differentiated in criterion-related validity studies as they relate differently with particular cognitive abilities. For example, verbal ability had its strongest relations with academic performance, although it was a negligible predictor for flying performance. Such results suggest being cautious when designing criterion-related validity studies and consider the criterion being used in the associations. The pattern noted for verbal ability in this study can be a good demonstrative example of the importance of such a call.

5.3 The Effects of Cognitive Abilities on Pilot Performance

The correlation and regression analyses used in many of AFOQT studies tested hypotheses about observed variables, whereas the SEM analyses used in this thesis tested hypotheses about latent constructs (e.g., Ullman & Bentler, 2012). While the results of regression analyses (e.g., hierarchal regression) are predictive in its essence, SEM analyses go beyond that to support explanatory relationship hypotheses (Pedhazur, 1997). A criticism often raised about the hierarchical regression analysis studies is that they are not appropriate for capturing latent cognitive constructs, and they do not provide simultaneous analysis for general and specific abilities (e.g., Oh et al., 2004; Glutting et al., 2006). Hence, the SEM bifactor models used in this study may overcome such criticism and enable simultaneous analysis of general and specific abilities. The three predictive studies revealed some important findings regarding the latent relationship between ability and performance in flying. Although the significant role of *g* in predicting job performance might not be questionable (Schmidt, 2002), the same may not be true for specific abilities. It has been argued that even a little incremental validity of specific abilities beyond that of *g* is seldom found and maybe not be attainable (Hunter, 1986). Hence, the findings of studies reported in this thesis represent a sought-after investigation to enrich the related literature and contribute to the debate of general versus specific abilities for predicting job performance.

As displayed in Table 20, across the four performance criteria, *g* effects were, on average, .38 on academic performance, .21 on primary phase performance, .15 on advanced phase performance, and .14 on overall performance. These estimates indicate an important prediction role of the general ability for pilot performance. After accounting for the domain-general factor underlying AFOQT subtests, there remains at least one or two domain-specific factors that uniquely predicted actual flying performance criteria, in the majority of analyses. Specifically, the average effects of aviation-related acquired knowledge on primary phase performance and overall performance were larger than those of *g* (.27 and .29 vs. .21 and .14, respectively) and likewise, the average effect of perceptual speed on overall performance was larger than that of *g* (.18 vs. .14).

Job knowledge tests often demonstrate strong relationships with job performance (Hunter, 1986; McDaniel et al., 1988). Hence, the relative importance of the aviation acquired knowledge factor in the current findings may be similar in the general population of flight organizations. Although it is common to hypothesize that job knowledge influences job performance indirectly through its relation with *g*, here, the direct effect of aviation-acquired knowledge appears to be greater than that of *g* (Ree et al., 1995; Schmidt et al., 1986). This result suggests a role for this broad ability factor that is unique from its overlap with *g*. Interestingly, knowledge-based test scores are also seen as indicators for the applicants' interest and motivation towards the job they are applying for (e.g., Kanfer & Ackerman, 1989; Colquitt et al., 2000), and thus, it may be this interaction between the cognitive and non-cognitive aspects of the construct that make this factor a robust predictor for pilot performance.

Verbal, quantitative, and spatial abilities appeared the least specific abilities incrementing the predictive validity beyond that gained through *g* measure. In most instances, the effects of these abilities on performance measures were either small, nonsignificant or even negatively significant, particularly verbal ability, suggesting a complex interplay between these specific abilities and *g* in predicting subsequent performance. Considering the significant correlations of quantitative and spatial abilities with most performance measures evidenced from the combined ability-performance correlated-factor models, the current SEM results suggest that these two particular abilities are core elements in the *g* score. Thus, the contribution of quantitative and spatial abilities to the predictive SEM models was fully accounted for by the general ability, leaving a small portion of variance to be explained uniquely by their specific constructs. The negative and insignificant estimates of verbal ability with most performance measures, either through the correlated-factor models or bifactor SEM models, clearly suggest that this ability has little to offer for the prediction of future pilot performance.

Thorndike (1985) suggested that the general ability score derived from a cognitive ability test battery explains nearly 85% to 90% of the predictable variance in criterion variables. Based on the current results, this conclusion might not be applicable to flying performance, as general ability did not show the expected strong relation with most criteria as compared to aviation-related acquired knowledge or even to perceptual speed in some samples and models. A large effect of *g* on general performance, with a few unique specific effects on specific performances, has been demonstrated in a number of studies (e.g., Gustafsson & Balke, 1993; McGrew et al., 1997; Keith, 1999). Due to the relatively small number of performance measures in each data set, it was not possible to consider modeling performance in this thesis in a design that can be more demonstrative for a general construct of performance and specific constructs of performance. Bifactor models, such as those considered for cognitive abilities here, are one potential configuration of latent performance modeling, whenever a sufficient number of performance measures are available. Despite this limitation in pursuing a method that provides a more accurate accounting of both general and specific constructs of performance, the studies of this thesis included several measures that arguably tap both breadths of performance. Specifically, three general flight performances (students' class rank in Sample 1 of the cross-validation study, the overall composite in Sample 2 of the cross-validation study, and the latent performance factor in the pilot sample of the cross-occupation study) and one general academic performance (the latent factor of academic achievement in Sample 3 of the cross-validation study) can be characterized as a general measures of performance. A common characteristics of these criteria is that they are all products of different dimensions and multiple measures of pilot performance, which can reasonably represent a general performance construct.

According to the ability–performance compatibility principle (Schneider & Newman, 2015), the general ability would be expected to be a stronger predictor of these particular general performance criteria than any other specific ability. Studies of this thesis, however, did not provide support for such an assumption, as the effects of general ability on the general flight performances were not distinctively different than the specific effects of two factors on these general measures. *g* effects varied from non-significant, such as that on students' rank criterion, to significant but lower than for the specific factor of aviation-related acquired knowledge, such as that on the observed overall performance composite and that on the latent performance factor (see Table 20). The differences in modeling the general performance could be one reason for this discrepancy between the relatively small effect of *g* on general flight performance in this thesis, and the large effect of *g* on general performance in studies that had a statistical operationalization of the general performance (e.g., bifactor). The nature of the flying job, as a highly technical job, can be an additional source of this unexpected pattern of general ability-general performance relationships.

The relation of general ability with general academic performance, nevertheless, had a different pattern, as seen in Sample 3 of the cross-validation study. Modeling 11 scores from tests of different academic flying topics as a latent factor showed that the general *g* factor was the main predictor for this overall academic performance ($\beta = .52$), while none of the specific factors (verbal, quantitative, acquired knowledge) contributed to prediction beyond the *g* factor. This is different from the conclusion about general flight performance, where the *g* score, along with two specific ability factors, accounted for a substantial proportion of its variance.

On the whole, results of bifactor predictive models showed that there is a need to go beyond *g* and consider specific abilities in order to understand the pilot's performance meaningfully. As emphasized by Brody (2002), *g* is not the only construct needed to describe individual differences. Current outcomes strongly indicate that *g*, acquired aviation knowledge, and perceptual speed should be the main focus when attempting to explain the three most crucial outcomes in flight performance: primary phase, advanced phase, and overall performance. Ability constructs of verbal, quantitative, and spatial abilities may be seen as only secondary variables that would not be expected to contribute uniquely to the predictive models beyond *g* score. Similarly, when trying to explain a pilot's academic performance in a training program, indeed, *g* appears to be the main contributor to performance. However, other specific constructs such as acquired aviation knowledge, quantitative ability, perceptual speed, and verbal ability can provide additional incremental validity to that produced by the *g* score. Spatial ability, in contrast, is not expected to increment the validity and thus, scores representing this construct need not be considered in predicting pilots' academic performance. On the whole, results reveal that at least two specific abilities are vital contributors to performance in flying.

5.4 Cognitive Abilities Across Three aviation Occupations

Through bifactor predictive models applied in the cross-occupation validation study, results clarified the interplay of general and specific cognitive abilities in predicting training performance in three aviation jobs: flying, navigation, and air battle management. The effect size of bifactor *g* was large in the navigator sample, small in the pilot sample, and negligible in the air battle manager sample. In contrast, the number of significant effects due to specific factors was none in the navigation sample, one in the flying sample, and three in the ABM sample. Due to the nature of the ABM performance measure as an average score of multiple written tests, the expectation was that this measure is more related to general ability than any specific ability as it is loaded with different academic and knowledge constructs. The influence of *g* on academic and achievement performance is a well-documented proposition (Gottfredson, 2002; Gustafsson & Undheim, 1996), especially when the performance is general in its scope (Kahana et al., 2002) as in this composite measure. However, contrary to expectations, quantitative ability, acquired aviation knowledge, and verbal ability were found to be better predictors of ABM performance than *g* after removing the general factor variance in their latent scores. Thus, the current findings that do not seem consistent with the majority of research, which frequently supports the dominant role of *g* over specific abilities in the prediction of academic performance, remain to be explained.

One possible reason for the significance of specific abilities and non-significance of *g* is the way that ABM performance was measured in this study as an observed variable indicated by one dimension of performance that related to academic achievement, rather than a latent variable comprising multiple indicators measuring different dimensions of performance. Including scores from multiple dimensions of ABM performance may make the construct more suitable to be predicted by a general predictor as *g* (e.g., Ones & Viswesvaran, 1996). Related to this, performance modeling of pilots and navigators in this study relied heavily on ratings of hands-on job samples, while that of ABM was mostly academic, which may not correspond well to the operationalization of *g* that includes spatial ability and perceptual speed, which are probably not required for conventional academic test items. Moreover, according to the job complexity hypothesis, a highly complex job requires more of general ability, and a less complex job requires only specific abilities (Gottfredson, 1997; Hunter et al., 1990; Murphy, 1989). Thus, the ABM performance in this study may have been represented by a less complex dimension in the wide criterion space of the ABM job, while the performance of pilots and navigators were represented by a global score with overlapped dimensions and constructs, most of which were practical in their essence. Furthermore, an ABM's job is generally less complex than pilot and navigator jobs (e.g., Fowley, 2016; Rhone, 2008), with a lower minimum qualifying scores (Carretta, 2008), and also based on the job complexity proposition, less role for *g* might be expected in this relatively less complex job in relation to other two aviation jobs. Last, it is expected that the courses taught in a technical program of training as ABM are also of technical scope and tend to target narrower knowledge and skills. According to the ability-criterion compatibility principle (Schneider & Newman, 2015), such a specific-oriented performance score is best predicted by a specific-oriented ability score.

In the navigator sample, when *g* was modeled, the effect of specific abilities either declined or faded away, as compared to their significant relationships with performance criteria in the correlated-factor model. *g* was found to be the only noteworthy predictor for navigators' performance, suggesting that the simple correlations of the five abilities with navigation performance were mostly due to their overlap with *g*. As for flying, navigation is a complex job that requires high cognitive ability. In the old 16-subtest AFOQT (e.g., Carretta, 1997), navigation applicants had to be qualified by 11 subtest composite scores (navigator/technical composite), as compared to 8 subtest composite scores (pilot composite) qualifying pilot applicants. This gives an indication of the cognitively demanding nature of this job that also may explain the greater role of *g*, relative to specific abilities, in the prediction of trainees' performance in navigation tasks.

The pattern in the pilots' sample comes in between these contrasting two patterns, where acquired aviation knowledge, along with *g*, stayed significant and effective in the predictive model. Acquired aviation knowledge became a better predictor of flight performance after removing the general factor variance from its scale scores. The effect of this factor was estimated to be .29

versus .11 for *g* factor. The higher effect of the aviation-related acquired knowledge factor in the pilot sample than in the other two samples may reflect the fact that the two indicators used to extract the factor are typically for content more related to pilot jobs than any other jobs in the USAF. The predictive utility of tests measuring acquired knowledge for pilot performance has been documented in a number of meta-analyses (ALMamari & Traynor, 2019; 2020; Hunter & Burke, 1994; Martinussen, 1996).

5.5 The Added Value of the Current Results

Current results about the predictive relations between AFOQT ability constructs and pilot performance extend our knowledge about complex relationships between abilities and job performance. Prior investigations of the AFOQT, such as those from which data were used in this thesis, provided useful information at the observed variable level and accumulated important findings about the utility of single ability tests or composite scores in predicting pilot performance (e.g., ALMamari & Traynor, 2020). Because a direct comparison between the current results and the results of primary studies from which correlation matrices were obtained is not possible due to the substantial differences in the goal of the study, analytic approach, and the modeling and specification of variables, this section focuses primarily on highlighting the added value of the current findings and the perceived advantages of applying a SEM approach that can supplement previous findings. For example, the primary validation study of this thesis added value to Johnson et al.'s (2017) study, supporting their conclusion about the incremental validity of perceptual speed subtest scores, but not spatial ability subtest scores, over AFOQT verbal, quantitative, and technical knowledge subtest scores for five performance measures. The result in this thesis suggest that the spatial ability construct has no incremental validity above *g* across three criteria, whereas the perceptual speed construct has a slightly lower effect than *g* on primary phase performance

(.23 vs. .26), slightly better effect than *g* on advanced phase performance (.16 vs. .12), and a much lower effect than *g* on academic performance (.06 vs. .24). This result is more revealing about the true effect of these two constructs on pilot performance than the .02 average joint incremental validity suggested by Johnson et al. (2017) for the SR and TR subtests, combined with five spatial ability subtests.

Sample 2 of the cross-validation study in this thesis provided a somewhat different conclusion than that reached by Olea and Ree's (1994) study. Conceptualizing *g* as the first component from principal component analysis of the AFOQT subtests and specific abilities as the remaining components (15 components/subtests) in Olea and Ree (1994) resulted in, via regression analysis, a conclusion that might have overestimated the role of *g* as a predictor for pilot performance, while underestimating the role of specific abilities in the prediction of future pilot training performance. Although this thesis supported a substantial role of *g*, it also demonstrates unique effects for aviation-related acquired knowledge and perceptual speed in predicting performance criteria. Moreover, the role of aviation-related acquired knowledge in this thesis was suggested to be even greater than that of *g* in predicting performance (.26 vs. .18 on average), and the role of perceptual speed was only slightly lower than that of *g* (.13 vs. .18 on average). This result was further confirmed in the pilot sample of the cross-occupation study that used the same data set but for other performance criteria, where modeling three performance indicators as a latent variable led to a similar conclusion regarding a substantial role of aviation-related acquired knowledge that even exceeded the role of *g* (.29 vs. .11).

Carretta and Ree (1997) tested a causal model for job acquisition of pilot job knowledge and flying skills, hypothesizing that prior job knowledge mediates the relationships between the general factor and subsequent training performance outcomes. Results suggested that the general ability has a direct effect on the acquisition of job knowledge, but indirect effect on the acquisition of flying skills. Sample 3 of the cross-validation study in this thesis had a different specification and organization of the variables than Carretta and Ree's (1997) study. Results revealed significant (and direct) effects of *g* on all three latent performance factors, including the two latent flying performance factors (β = .26 on primary phase performance and .30 on advanced phase performance). Given the line of research of Ree and his colleagues about the primacy role of a general factor in job performance over specific abilities (e.g., Ree & Earles, 1991; Ree & Earles, 1996; Ree et al., 1994), and more recently, their assumption about the pervasiveness of dominant general factors in organizational measurement (Ree et al., 2015), the results from the cross-validation study of Sample 3 provide support for a *direct* (and large) effect of *g* on the two flying latent performance factors, as compared to the *indirect* effect of *g* on the same factors and nonsignificant *direct* effect suggested by Carretta and Ree (1994).

The simple correlation analysis applied in Carretta and Ree (1995) and Duke and Ree (1996) was useful in identifying AFOQT subtests that are most predictive of pilot performance measures. However, a large number of correlation coefficients and the limitation of inferences obtained from single ability tests may limit the generalizability of results and restrict extending the interpretations to a more abstract level (e.g., theory). Results in this thesis using the same two data sets (primary validation study and Sample 1 of cross-validation study) provided a more focused conclusion, suggesting that the source of AFOQT validity for pilot performance measures can be explained by one to three main constructs: aviation-related acquired knowledge, perceptual speed, and the general factor of ability. Conversely, the unique effects of verbal, quantitative, and spatial abilities were found to be trivial in the studies of this thesis, after accounting for the *g* effect, indicating that the predictive power exhibited by subtests indicative of these abilities are mostly due to their *g*-loadings.

5.6 Implications and Recommendations

The main aim in this thesis was to address the lack of research evidence on the predictive relations between cognitive abilities and pilot performance outcomes. I have done so by reanalyzing several data sets using SEM approaches, with special attention to bifactor modeling, whereby the effect on criteria due to the general factor is disentangled from the effects due to the specific ability factors. Accordingly, results of this thesis can be informative with useful implications for psychometricians, AFOQT developers, practitioners, researchers in aviation psychology, administrators of university flight programs, and designers of youth programs targeting future pilots.

For psychometricians developing or maintaining selection test batteries, it is recommended that they increase the attention given to the construct of aviation acquired knowledge in the pilot selection process as it is shown to be the best predicting factor of several pilot performances. The second construct that should be emphasized in the selection process is perceptual speed due to its uniqueness in predicting performance. For the general factor of ability, its significance as a predictor for flight performance is certainly evidenced and it should remain a supportive predictor in the selection of pilots, navigators, or even air battle managers. Conversely, verbal ability as a predictor for flight performance is questionable and may require further attention. The current analysis indicated that scores of verbal ability failed to show any noteworthy effect, either in correlated-factor models or bifactor SEM models. If studies continued to show a lack of association between scores of verbal ability and flight performance, a suggestion could be made to disregard this ability in test batteries that are specifically designed for pilot selection.

Presently, the AFOQT developers compute six composite scores from ten AFOQT subtests (Form T). The construction of these composites is mostly conceptual groupings of the subtests, rather than factor-based. As emphasized by Glutting et al. (2006), factor scores have several strengths over conceptual subtest groupings or single test scores, including (1) stronger construct validity, (2) higher reliability, and (3) better representing phenomena by separating subtest specificity, method variance, and measurement error. The current thesis highlighted potential promise for latent variable modeling approaches to cognitive ability and job performance relations. In a recent study, Benson et al. (2016) argued that orthogonal refined factor scores (i.e., construct scores derived from a bifactor model) can be a better representation of the associations between CHC broad ability constructs and achievement outcomes than non-refined scores (i.e., factor scores created by summing subtest scores) do; thus, calling for more application of scores based on latent variable models. Hence, Future scoring development of the AFOQT may include factor scores to overcome problems typically associated with the more common practice of subtest analysis.

Results in this thesis suggest that aviation practitioners should continue considering specific abilities and go beyond *g* to understand the pilot's latent cognitive abilities meaningfully. At the same time, the view of cognitive constructs, at least those extracted from the AFOQT, should not be of equal weight. Some abilities demonstrated stronger relations with pilot performance measures than others. For example, when attempting to predict the pilot's hands-on flying performance using AFOQT scores, more focus should be given to the three constructs emphasized above: *g*, aviation acquired knowledge and perceptual speed. However, when explaining a pilot's academic performance, the focus should be centered, for the most part, on *g*, although other constructs such as acquired aviation knowledge, quantitative ability, perceptual speed, and verbal ability will be expected to increment the validity marginally.

Researchers in aviation psychology should be thoughtful about the design of criterion-related validity studies, and the suitability of particular performance criteria, e.g., a general ability would be expected to perform better with an overall academic grade as a criterion, while aviation-related acquired knowledge would be expected to perform better with a criterion of actual flying performance. Moreover, the current findings highlight the importance of assessing predictive relations with the SEM approach to provide evidence justifying the continued and expanded use of particular cognitive ability tests for pilot selection. It extends our understanding of the impact of the broad ability constructs on the prediction of pilot performance across different criteria. It is recommended that researchers make use of the methods applied in this thesis and attempt to replicate the results, perhaps using different cognitive ability tests, pilot performance measures, or flying organization settings. It can serve as a potential framework for future research on validity evidence of ability factor scores derived from test batteries.

Another recommendation is addressed to administrators of university flight programs. It is common for universities to admit students to flight programs based on educational criteria such as GPA, SAT scores, or ACT scores with no measurement of specific cognitive abilities that are relevant to flying performance. The findings in this study provide useful evidence for universities to consider introducing more relevant screening tests as criteria for flight program admission. Given the support provided by this thesis for a critical role for the general ability in flying, the current conventional admission criteria (e.g., GPA, ACT) may cover this aspect of predictive models to a reasonable extent. However, there will be a need for specific tests measuring other essential aspects of the predictive models such as aviation-related acquired knowledge and perceptual speed ability. These particular two constructs are highly recommended for consideration in any admission process targeting pilot applicants. The subtests of Instrument Comprehension, Aviation Knowledge, Table Reading, and Scale Reading are examples of tests suitable for capturing these two constructs. All of the four tests are of paper-and-pencil type, which motivates the incorporation of similar tests in the selection process.

Finally, specific recommendations from a selection perspective for youth program administrators targeting flying as a potential career are to (a) emphasize the importance of cognitive abilities as selection criteria for flight programs and that these abilities can be developed by practice, (b) determine specific cognitive abilities that serve as significant predictors of pilot performance among young people, (c) familiarize participants with common ability test batteries used in pilot selection, and (d) provide opportunities for training directed to the most wanted cognitive abilities for pilots.

5.7 Limitations and Future Research

Despite the useful theoretical and practical implications of this thesis, there are some limitations to be noted when interpreting the results, which can also be informative for future research on ability-performance relationships. First, in spite of the different performance measures (13 total pilot performance indexes) and modeling strategies (i.e., latent, observed) used across analyses, the studies reported in this thesis showed an acceptable level of consistency in the relative importance of g and specific abilities as predictors for pilot performance. However, the magnitude of structural regression coefficients varied notably for each construct. One clear explanation for this variation is the difference in the type of performance measures and the phase of training from which they were collected. Future cross-validation studies may attempt to obtain performance data that are similar in content and method of collection, perhaps similar to that presented by Carretta et al. (2014). Academic performance should always be differentiated from hands-on flying performance when planning criterion-related validity investigations due to their distinct patterns of association with cognitive abilities.

Second, the findings of the current thesis can be generalized to a good extent. Given the meta-analytic approach used in the preliminary study, the multiple pilot samples (4 samples) used in the primary validation and cross-validation studies, the across-occupation comparison attempted in the cross-occupation validation study, and more importantly, the SEM approach used in the analyses, the resulting findings can be taken with reasonable level of confidence. However, the data of this thesis are primarily from the population of USAF undergraduate pilot students in their flight training program. Despite the notable strength of using multiple similar samples, such as reducing the likely effect of sample-specific differences on the results, the findings may be less generalizable. Similarly, cognitive data assessed in the thesis were all sourced from one cognitive test battery, the AFOQT. The use of the same test battery facilitate a direct comparison across samples; however, conclusions drawn from such a battery may be less generalizable. Despite the fact that analyzing several data sets in this thesis minimizes these two limitations, it is recommended that future research expand the scope of investigations to include different ability tests obtained from different flying organization settings to provide further evidence for constructs' predictive relations with flight performance measures. Relatedly, considering that the AFOQT is a general selection instrument for all USAF officer candidates, not exclusive for aircrew selection, and that another test battery specifically designed for pilot candidates also exists (i.e., TBAS; Carretta, 2011), it can be expected that the subtests incorporated in the AFOQT are those with potential to be predictive of several USAF occupations, not only of flying jobs. Thus, it is also recommended that future studies focus on ability constructs that have shown to be crucial for pilot performance, but not measured presently by the AFOQT. Examples of such constructs include working memory (Wang et al., 2018), reaction time (King et al., 2013), and multitasking (Barron & Rose, 2017). Adding some of these constructs as predictors in pilot performance studies can be essential.

Third, despite the multiple criteria used in this thesis, future studies may expand the scope of investigation of pilot performance measures to include scores and ratings covering more specific areas of flight training (e.g., Bates et al., 1997). For instance, Sample 1 of the cross-validation study in this thesis attempted a somewhat different criterion, using pilot's exceedance of average flying hours as a performance measure. The conclusion determined for the effect of *g* on this specific criterion was different from those determined with other performance criteria in this thesis. It is also suggested that some well-accepted performance models be used in criterion-related validity studies to explore the utility of intelligence tests in predicting performance dimensions in the pilot job that have not been explored in previous studies. For example, Campbell and Wiernik (2015) proposed an eight-factor performance model that accommodates some critical dimensions of individual work performance. Technical performance, the most assessed dimension in pilot performance research, is only one dimension among many others that can be attended to in criterion-related validity studies. Barron, Carretta, and Bonto-Kane (2016) and Barron, Carretta, and Rose (2016) provided examples for a holistic approach to pilot performance by evaluating officer performance reports. Exploring different performance outcomes may produce different predictive relations with cognitive abilities. Nonetheless, the performance measures targeted in this thesis remain the most common performance criteria in flying, and hence, they are representative of typical pilot performance.

Fourth, the sample sizes of data sets utilized in the current thesis were generally large for whole-group analyses. However, they were not divided by, for example, gender, age or ethnicity group so as to support subgroup analysis for a better understanding of the results. One dataset in the cross-validation study (Sample 3) utilized male pilots' data from a technical report that also included female pilots' data, and the initial plan was to make multigroup SEM to detect differences in the predictive relations as a function of pilot gender. However, the unbalanced size of the two samples (2% for female sample) did not allow for such a research plan. Future research may focus on the measurement invariance to give further evidence for the cognitive abilities–pilot performance predictive relationships and whether they differ among, for example, gender and ethnicity membership. Fifth, in spite of the several datasets utilized in this thesis and the replicability of many results, the modeling techniques were based primarily on a bifactor model, which has some inherent limitation (Reise, 2012; Reise et al., 2010). For example, the factors in a bifactor model, either general factor or grouping factors, are restricted to be uncorrelated. Also, each indicator in a bifactor model is allowed to load onto the general factor, and to only one grouping factor. Due to the known intercorrelation between cognitive data, these assumptions may seem unrealistic in some occasions, where group factors are conceptually related, or an indicator can mark more than one construct. For example, perceptual speed and spatial ability factors are expected to share some common variance that is attributable to the general factor (e.g., Barron & Rose, 2013). Similarly, the Scale Reading subtest has salient loadings ($\lambda > .30$) on perceptual speed and quantitative ability factors, although it was modeled as a marker only for the perceptual speed construct. Thus, it would be useful to attempt different approaches with other analytic procedures to give the results further credibility. Example of such approaches that have recently gained popularity includes relative weight analysis (Johnson, 2000), dominance analysis (Azen &

Budescu, 2003), and higher-order models that specify the general factor and all first-order factors as predictors, and correlate the error of specific factors with performance outcomes (Coyle, 2018). Replicating the results of the current thesis using some of these methods can give further confidence in the results.

Sixth, another limitation is that some ability factors were represented by as few indicators as two, which may be inadequate in capturing the construct under investigation. Three or more indicators are typically recommended (Kenny et al., 1998; Marsh et al., 1998). However, other researchers found that two or even one indicator may be sufficient (Hayduk & Littvay, 2012). A large sample size, like those used in this thesis, can also help to compensate for a few indicators per factor (Koran, 2020). Additionally, the several factor-analytic assessments of AFOQT data attempted in this thesis, either via EFA or CFA, provided evidence supporting the construct validity of the ability factors, even with the relatively fewer indicators measuring their constructs. Interestingly, aviation-related acquired knowledge, the strongest predictor of pilot performance revealed in this thesis, has only two indicators, Instrument Comprehension and Aviation Information. The second strongest predictor was perceptual speed, which also indicated by two subtests: Table Reading and Scale Reading. Therefore, the influence of the number of indicators seems minor in this thesis.

Seventh, the results of studies in this thesis were a function of two types of variables, independent (i.e., predictor) and dependent (i.e., criterion) variables, of which their relations were modeled using direct paths from predictors to criterion. No moderation nor mediation variables that have the potential to influence the predictor-criterion relationship were considered. Nonintellectual factors such as socioeconomic status, cultural familiarity, test-taking strategies, or reading ability may have some effects on the findings. Additionally, researchers started enquiring whether pilot students who are enrolled in aviation programs tend to exhibit higher rates of depression, stress, or anxiety (e.g., Allsop & Gray, 2014; Jacobs et al., 2020; Sloan et al., 2018), and whether the use of certain stress coping skills can lower stress level of trainees and reduce their likelihood of errors (e.g., Kirschner, 2011). Including some of these variables in the models may influence the results. Future research should consider adding potential, influential covariates, or methods to understand predictive relations more soundly. Adding some of these variables may improve the power of ability-performance predictive models.

Eight, the significant predictors of pilot performance found in this thesis yielded generally small structural regression coefficients. Due to the range restriction, studies involving job selection, by nature, limit the range of observed variability in a sample. The results, therefore, are likely underestimates of the true relations between cognitive abilities and pilot performance. This limitation is well-known in the literature and is often treated with some forms of correction (e.g., Raju & Brand, 2003; Sackett & Yang, 2000). Correction of range restriction, as well as measurement unreliability, typically yield stronger true relationships between cognitive predictors and performance criteria. In this study, I used the uncorrected data, not the corrected data, to provide results that can be more matching to the real-world causal structure as suggested by cognitive testing and pilot performance outcomes. The weakness in the observed correlations requires reconsideration of the interpretation of effect sizes, taking into account the norms in the field. The recent study of Gignac and Szodorai (2016) showed that lower cut-off scores for interpreting the criterion-related validity in organizational psychology might be more realistic than those provided by some scholars (Cohen, 1988; Lipsey & Wilson, 2001). In aviation, even these newly recommended cut-off scores can be higher than the normative data of correlations reported in the field. Thus, future research may focus on the validation studies in aviation to suggest values that can be more representative than those interpreted in other studies.

Ninth, the AFOQT is developed and maintained by a research team from the USAF, and most published and unpublished AFOQT studies are authored by members of this team. All studies included in the meta-analytic investigation of the preliminary study were conducted by the same group of researchers and also, at somewhat close period of time, which may raise a question about the dependency of the synthesized correlation coefficients. One critical assumption in meta-analysis is that the effect sizes being synthesized are independent, and failure to meet this assumption could cause misleading or even wrong results (Cheung, 2019). Hence, this observation about studies included in the meta-analysis needs to be borne in mind when interpreting the results of the preliminary study. Finally, some validation investigations have used contemporary intelligence models of human cognitive structure as a framework. Due to the practical organizational goals of AFOQT use, its coverage does not correspond to any particular intelligence framework. Future research may attempt to pursue one of the modern intelligence models to expand the cognitive ability constructs investigated and advance the understanding of constructs that might be missing from current flight selection batteries. This will also allow a better comparison with the research conducted in other domains (e.g., educational, neuropsychological).

5.8 Conclusion

In view of the limited number of SEM studies in aviation, our understanding of ability-performance relationships has relied mostly on observed scores of single ability tests or composites derived from multiple ability tests. The findings presented here are useful because they provide validity assessment for five latent constructs of specific abilities, as well as the general factor of ability, using large pilot and non-pilot samples via two latent modeling approaches of predictive ability (i.e., ability-performance correlated-factor and bifactor SEM models). On the basis of the correlation-based validity approach, results indicated that quantitative, spatial, perceptual speed, and acquired aviation knowledge are generally valid predictors for pilot performance, while verbal ability did not relate significantly with pilot performance measures, except with academic performance. On the contrary, after accounting for the effect of the general ability, the validity approach based on the bifactor SEM model revealed that only acquired aviation knowledge and perceptual speed, in addition to *g*, have predictive power for most flying performance measures. This finding largely supports the utility of cognitive ability test scores in pilot selection, but with varying degrees of predictive power and effect sizes. Informed by the current results, practitioners and researchers in aviation psychology may gain further insight about the cognitive abilities' contribution to successful training and flying performance. Flying organizations (e.g., military, airlines, university programs) may utilize some of the present findings to improve their selection test batteries in their continued attempt to select the right applicants and reduce the attrition rates in flight programs. The conclusions reached can serve as a basis for further investigations of the relationship between pilot applicants' measured ability and their subsequent training performance.

REFERENCES

References marked with an asterisk (*) indicate studies included in the meta- analysis

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*(3), 288-313.

Ackerman, P. L., & Beier, M. E. (2012). The problem is in the definition: g and intelligence in I–O psychology. *Industrial and Organizational Psychology*, *5*(2), 149-153.

Ackerman, P. L., Kanfer, R., & Goff, M. (1995). Cognitive and noncognitive determinants and consequences of complex skill acquisition. *Journal of Experimental Psychology: Applied*, *1*(4), 270-304.

Aguilar, I. D. (2017). *Comparison of the Air Force Officer Qualifying Test Form T and Form S: Initial item-and subtest-level analyses* (No. AFCAPS- TR-2017-0002). Randolph AFB, San Antonio, TX: Air Force Personnel Center.

Agnello, P., Ryan, R., & Yusko, K. P. (2015). Implications of modern intelligence research for assessing intelligence in the workplace. *Human Resource Management Review*, *25*(1), 47-55.

ALMamari, K., & Traynor, A. (2019). Multiple test batteries as predictors for pilot performance: A meta-analytic investigation. *International Journal of Selection and Assessment*, *27*(4), 337-356.

ALMamari, K., & Traynor, A. (2020). Predictive Validity of the Air Force Officer Qualifying Test (AFOQT) for Pilot Performance: A Meta-analytic Investigation at the Subtest Level. *Aviation Psychology and Applied Human Factors*.

Allsop, J., & Gray, R. (2014). Flying under pressure: Effects of anxiety on attention and gaze behavior in aviation. *Journal of Applied Research in Memory and Cognition*, *3*(2), 63-71.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411-423.

Arth, T. O. (1986). *Validation of the AFOQT (Air Force Officer Qualifying Test) for Non-Rated Officers* (No. AFHRL-TP-85-50). Brooks Air Force Base, San Antonio, TX: Air Force Human Resources Laboratory.

*Arth, T. O., Steuck, K. W., Sorrentino, C. T., & Burke, E. F. (1990). *Air Force Officer Qualifying Test (AFOQT): Predictors of Undergraduate Pilot Training and Undergraduate Navigator Training Success* (No. AFHRL-TP-89-52). Brooks Air Force Base, San Antonio, TX: Air Force Human Resources Laboratory.

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, *8*(2), 129-148.

Bacharach, S.B. (1989). Organizational theories: some criteria for evaluation. *The Academy of Management Review*, *14*(4), 496–516.

Bagozzi, R. P., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, *40*(1), 8-34.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*(1), 78-102.

Barron, L. G., Carretta, T. R., & Bonto-Kane, M. V. A. (2016). Relations of Personality Traits to Military Aviator Performance: It Depends on the Criterion. *Aviation Psychology and Applied Human Factors*, *6*(2), 57–67.

Barron, L. G., Carretta, T. R., & Rose, M. R. (2016). Aptitude and trait predictors of manned and unmanned aircraft pilot job performance. *Military Psychology*, *28*(2), 65-77.

Barron, L. G., & Rose, M. R. (2017). Multitasking as a predictor of pilot performance: Validity beyond serial single-task assessments. *Military Psychology*, *29*(4), 316-326.

Bartram, D., & Baxter, P. (1996). Validation of the Cathay Pacific Airways pilot selection program. *The International Journal of Aviation Psychology*, *6*(2), 149-169.

Bates, M. J., Colwell, C. D., King, R. E., Siem, F. M., & Zelenski, W. E. (1997). *Pilot performance variables* (No. AL/CF-TR-1997-0059). Brooks AFB, TX: Armstrong Laboratory, Manpower and Personnel Division.

Beaujean, A. A. (2015). John Carroll’s views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence*, *3*(4), 121-136.

Beier, M. E., Kell, H. J., & Lang, J. W. (2019). Commenting on the “Great Debate”: General Abilities, Specific Abilities, and the Tools of the Trade. *Journal of Intelligence*, *7*(1), 85-95.

Benson, N. F., Kranzler, J. H., & Floyd, R. G. (2016). Examining the integrity of measurement of cognitive abilities in the prediction of achievement: Comparisons and contrasts across variables from higher-order and bifactor models. *Journal of School Psychology*, *58*, 1-19.

Bentler, P. M. (1995). *EQS structural equations program manual* (Vol. 6). Encino, CA: Multivariate software.

Berger, F. R., Gupta, W. B., Berger, R. M., & Skinner, J. (1990). *Air Force Officer Qualifying Test (AFOQT) form P manual*. Tech. Rep. No. AFHRL-TR-89-56). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Blasco-Belled, A., Rogoza, R., Torrelles-Nadal, C., & Alsinet, C. (2019). Emotional intelligence structure and its relationship with life satisfaction and happiness: New findings from the bifactor model. *Journal of Happiness Studies*, 1-19.

Brody, N. (2002). g and the one–many problem: Is one enough? In G. R. Bock, J. A. Goode, & K. Webb, *The nature of intelligence: Novartis Foundation Symposium 233* (pp. 122–135). New York, NY: Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230-258.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796-846.

Burgoyne, A. P., Harris, L. J., & Hambrick, D. Z. (2019). Predicting piano skill acquisition in beginners: The role of general intelligence, music aptitude, and mindset. *Intelligence*, *76*, 101383.

Burke, E., Hobson, C., & Linsky, C. (1997). Large sample validations of three general predictors of pilot training success. *The International Journal of Aviation Psychology*, *7*(3), 225-234.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge.

Callender, M. N. (2018). Transfer and Cost Effectiveness as Guides to Simulator/Flight Training Device Use. *The Collegiate Aviation Review International*, *26*(1), 28-32.

Campbell, J. S., Castaneda, M., & Pulos, S. (2009). Meta-analysis of personality assessments as predictors of military aviation training success. *The International Journal of Aviation Psychology*, *20*(1), 92-109.

Campbell, D.T., & Stanley, J.C. (1966). Experimental and quasi-experimental designs for research. In N.L. Gage, *Handbook of research on teaching* (pp. 1-76). Chicago, IL: Rand-McNally.

Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual Review of Organizational Psychology and Organizational Behavior*, *(2)*1, 47-74

Caponecchia, C., Zheng, W. Y., & Regan, M. A. (2018). Selecting trainee pilots: Predictive validity of the WOMBAT situational awareness pilot selection test. *Applied Ergonomics*, *73*, 100-107.

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements.* Gottingen, Germany: Hogrefe.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In J. L. Genshaft & D. P. Flanagan (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (Vol. xvi, pp. 122–130). New York: Guilford Press.

Carretta, T. R. (2000). US Air Force pilot selection and training methods. *Aviation, Space, and Environmental Medicine*, *71*(9), 950-956.

Carretta, T. R. (2005). *Development and validation of the Test of Basic Aviation Skills (TBAS)*. Wright-Patterson, OH: Air Force Research Laboratory, Human Effectiveness Directorate.

Carretta, T. R. (2008). *Predictive validity of the Air Force Officer Qualifying Test for USAF air battle manager training performance* (No. AFRL-RH-WP-TR-2009-0007). Wright-Patterson AFB, OH: Air Force Research Laboratory, Human Effectiveness Directorate.

Carretta, T. R. (2010). Predictive validity of the Air Force Officer Qualifying Test for non-rated officer specialties. *Military Psychology*, *22*(4), 450-464.

Carretta, T. R. (2011). Pilot candidate selection method. *Aviation Psychology and Applied Human Factors*, *1*(1), 3-8.

*Carretta, T. R., & Ree, M. J. (1995). Air Force Officer Qualifying Test validity for predicting pilot training performance. *Journal of Business and Psychology*, *9*(4), 379-388.

Carretta, T. R., & Ree, M. J. (1996). Factor structure of the air force officer qualifying test: Analysis and comparison. *Military Psychology*, *8*(1), 29-42.

Carretta, T. R., & Ree, M. J. (1997). A preliminary evaluation of causal models of male and female acquisition of pilot skills. *The International Journal of Aviation Psychology*, *7*(4), 353-364.

Carretta, T. R., & Ree, M. J. (2003). Pilot selection methods. In P. S. Tsang & M. A. Vidulich (Eds.), *Human factors in transportation: Principles and practice of aviation psychology* (pp. 357–396). Mahwah, NJ: Erlbaum.

*Carretta, T. R., Retzlaff, P. D., & King, R. E. (1997). *A Tale of Two Test Batteries: A Comparison of the Air Force Officer Qualifying Test and the Multidimensional Aptitude Battery* (No. AL/HR-TP-1997-0052). Williams AFB, AZ: Armstrong Laboratory, Human Resources Directorate. Carretta, T. R., Rodgers, M. N., & Hansen, I. (1996). The Identification of Ability Requirements and Selection Instruments for Fast Jet Pilot Training. *Euro-NATO ACHFWG Technical Report-2*.

Carretta, T. R., Rose, M. R., & Trent, J. D. (2016). *Air Force Officer Qualifying Test Form T: Initial Item-, Test-, Factor-, and Composite-Level Analyses*. Wright-Patterson AFB, OH: Air Force Research Laboratory, Human Performance Wing.

Carretta, T. R., Teachout, M. S., Ree, M. J., Barto, E. L., King, R. E., & Michaels, C. F. (2014). Consistency of the relations of cognitive ability and personality traits to pilot training performance. *The International Journal of Aviation Psychology*, *24*(4), 247-264.

Carretta, T. R., Zelenski, W. E., & Ree, M. J. (2000). Basic Attributes Test (BAT) retest performance. *Military Psychology*, *12*(3), 221-232.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1-22.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Causse, M., Dehais, F., & Pastor, J. (2011). Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots. *The International Journal of Aviation Psychology*, *21*(3), 217-234.

Cheung, M. W. L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*, 1521.

Cheung, M. W. L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology review*, 1-10.

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: a two-stage approach. *Psychological Methods*, *10*(1), 40-64.

Cheung, M. W., & Chan, W. (2009). A two-stage approach to synthesizing covariance matrices in meta-analytic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(1), 28-53.

Cheung, M. W.-L., & Cheung, S. F. (2016). Random-effects models for meta-analytic structural equation modeling: Review, issues, and illustrations. *Research Synthesis Methods*, *7*, 140–155.

Child, D. (2006). *The essentials of factor analysis* (3rd ed.). New York: Continuum.

Christensen, A. P., Silvia, P. J., Nusbaum, E. C., & Beaty, R. E. (2018). Clever people: Intelligence and humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 136-143.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: a meta-analytic path analysis of 20 years of research. *Journal of applied psychology*, *85*(5), 678-707.

Coyle, T. R. (2014). Predictive validity of non-g residuals of tests: More than g. *Journal of Intelligence*, *2*(1), 21-25.

Coyle, T. R. (2018). Non-g factors predict educational and occupational criteria: More than g. *Journal of Intelligence*, *6*(3), 43.

Coyle, T. R., & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing g. *Intelligence*, *36*(6), 719-729.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Cucina, J., & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence*, *5*(3), 27.

Damitz, M., Manzey, D., Kleinmann, M., & Severin, K. (2003). Assessment center for pilot selection: Construct and criterion validity and the impact of assessor type. *Applied Psychology*, *52*(2), 193-212.

Damos, D. L. (1993). Using meta-analysis to compare the predictive validity of single-and multiple-task measures to flight performance. *Human Factors*, *35*(4), 615-628.

Damos, D. (2011). KSAOs for military pilot selection: A review of the literature (AFCAPS-FR2011– 0003). Randolph AFB, TX: Air Force Personnel Center, Strategic Research and Assessment Branch.

Deary, I. J., Irwing, P., Der, G., & Bates, T. C. (2007). Brother–sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence, 35*(5), 451-456.

De Winter, J. C., Dodou, D., & Mulder, M. (2012). Training effectiveness of whole body flight simulator motion: A comprehensive meta-analysis. *The International Journal of Aviation Psychology*, *22*(2), 164-183.

Dombrowski, S. C. (2015). Exploratory bifactor analysis of the WJ-III Achievement at School Age via the Schmid–Leiman orthogonalization procedure. *Canadian Journal of School Psychology*, *30*(1), 34-50.

Drasgow, F. (2012). Intelligence and the workplace. In I. B. Weiner, N. W. Schmitt, & S. Highouse (Eds.), *Handbook of psychology, industrial and organizational psychology*. London, England: Wiley.

Drasgow, F., Nye, C. D., Carretta, T. R., & Ree, M. J. (2010). Factor structure of the Air Force Officer Qualifying Test form S: Analysis and comparison with previous forms. *Military Psychology*, *22*(1), 68-85.

Driskell, J. E., & Adams, R. J. (1992). *Crew resource management: An introductory handbook*. Washington, DC: U.S. Department of Transportation, Federal Aviation Administration, Research Development Service.

Driskell, J. E., & Olmstead, B. (1989). Psychology and the military: Research applications and trends. *American Psychologist*, *44*(1), 43-54.

*Duke, A. P., & Ree, M. J. (1996). Better candidates fly fewer training hours: Another time testing pays off. *International Journal of Selection and Assessment*, *4*(3), 115-121.

Earles, J. A., & Ree, M. J. (1991). *Air Force Officer Qualifying Test (AFOQT): Estimating the General Ability Component* (No. AL-TP-1991-0039). Brooks AFB, TX: Armstrong Laboratory.

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence*, *6*(3), 9-31.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*(4), 617-634.

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY: Guilford Press.

Evermann, J., & Tate, M. (2009). Building theory from quantitative studies, or, how to fit SEM models. *Proceedings of the International Conference on Information Systems (ICIS)*, paper 192.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272.

Finegold, L. S., & Rogers, D. (1985). *Relationship between Air Force Officer Qualifying Test scores and success in air weapons controller training* (No. AFHRL-TR-85-13). Brooks AFB, TX: Air Force Human Resources Laboratory.

Flather, M. D., Farkouh, M. E., Pogue, J. M., & Yusuf, S. (1997). Strengths and limitations of meta-analysis: larger studies may be more reliable. *Controlled clinical trials*, *18*(6), 568-579.

Fleishman, E. A. (1956). Psychomotor selection tests: Research and application in the United States Air Force. *Personnel Psychology*, *9*(4), 449-467.

Fleishman, E.A. (1992). *The Fleishman Job Analysis Survey (F-JAS)*. Palo Alto: Consulting Psychologists Press, Inc.

Floyd, R. G., Shands, E. I., Rafael, F. A., Bergeron, R., & McGrew, K. S. (2009). The dependability of general-factor loadings: The effects of factor-extraction methods, test battery composition, test battery size, and their interactions. *Intelligence*, *37*(5), 453-465.

Fowley, J. W. (2016). *Undergraduate Air Battle Manager Training: Prepared to Achieve Combat Mission Ready*. Maxwell AFB, Alabama: Air Command and Staff College, Distance Learning, Air University United States.

Ganzach, Y., & Patel, P. C. (2018). Wages, mental abilities and assessments in large scale international surveys: Still not much more than g. *Intelligence*, *69*, 1-7.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement*, *71*(3), 551-570.

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor?. *Psychology Science*, *50*(1), 21-43.

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence*, *62*, 138-147.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74-78.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, *12*(6), 1-28.

Glomb, T. M., Earles, J.A. (1997). *Air Force Officer Qualifying Test (AFOQT): Forms Q development, preliminary equating and operational equating* (No. AL/HR-TP-1996-0036). Brooks AFB, San Antonio, TX: Human Resources Directorate, Armstrong Laboratory.

Glutting, J. J., Watkins, M. W., Konold, T. R., & McDermott, P. A. (2006). Distinctions without a difference: The utility of observed versus latent factors from the WISC-IV in estimating reading and math achievement on the WIAT-II. *The Journal of Special Education*, *40*(2), 103-114.

Goertz, W., Hülsheger, U. R., & Maier, G. W. (2014). The validity of specific cognitive abilities for the prediction of training success in Germany: A meta-analysis. *Journal of Personnel Psychology*, *13*(3), 123-133.

Goeters, K. M., & Maschke, P. (2004). Cost-benefit analysis of pilot selection: The economic value of psychological testing. In *Aviation psychology: Practice and research*, ed. K. M. Goeters, 203-208. Aldershot: Ashgate Publishing Limited

Goeters, K. M., Maschke, P., & Eißfeldt, H. (2004). Ability requirements in core aviation professions: Job analyses of airline pilots and air traffic controllers. In *Aviation psychology: Practice and research*, ed. K. M. Goeters, 99-119. Aldershot: Ashgate Publishing Limited.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79-132.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, *15*(1-2), 25-46.

Griffin, G. R., & Koonce, J. M. (1996). Review of psychomotor skills in pilot selection research of the US military services. *The International Journal of Aviation Psychology*, *6*(2), 125-147.

Guo, B., Perron, B. E., & Gillespie, D. F. (2009). A systematic review of structural equation modeling in social work research. *British Journal of Social Work*, *39*(8), 1556-1574.

Gustafsson, J-E. (2001). On the hierarchical structure of personality and ability. In J. Collis & S. Messick (Eds), Intelligence and personality: Bridging the gap in theory and measurement. Mahwah, NJ: Lawrence Erlbaum Associates.

Gustafsson, J.-E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73–95). Hillsdale, NJ: Erlbaum.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407-434.

Gustafsson, J. E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D.C. Berliner, R.C. Calfee (Eds.), *Handbook of educational psychology* (pp. 186-242). New York: Prentice Hall International.

Hampton, S., Truong, D., Byrnes, K., & Techau, T. (2017). Pilot Training Metrics at a Part 141 University Training Program. *17th AIAA Aviation Technology, Integration, and Operations Conference*, 3087.

Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology*, *12*(1), 159.

Hays, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military Psychology, 4*(2), 63-47.

Hedge, J. W., Bruskiewicz, K. T., Borman, W. C., Hanson, M. A., Logan, K. K., & Siem, F. M. (2000). Selecting pilots with crew resource management skills. *The International Journal of Aviation Psychology*, *10*(4), 377-392.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558.

Hoermann, H., & Damos, D. L. (2019). The Use of a Perceptual Speed Test in Civilian Pilot Selection. *20th International Symposium on Aviation Psychology*, 391-396.

Hoermann, H. J., & Goerke, P. (2014). Assessment of social competence for pilot selection. *The International Journal of Aviation Psychology*, *24*(1), 6-28.

Hoermann, H-J. & Guan, H.J. (2002). Development and evaluation of psychometric selection methods for Chinese ab-initio student pilots. *Chinese Journal of Aerospace Medicine*, 13(2), 102.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41-54.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53-60.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review*, *75*(3), 242-259.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *Woodcock-Johnson technical manual* (pp. 197–232). Chicago, IL: Riverside.

Howse, W. R. (2011). *Knowledge, skills, abilities, and other characteristics for remotely piloted aircraft pilots and operators* (AFCAPS-FR-2011– 0006). Randolph AFB, TX: Air Force Personnel Center, Strategic Research and Assessment Branch.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology*, *13*(4), 373-386.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, *29*(3), 340-362.

Hunter, D. R., & Burke, E. F. (1994). Predicting aircraft pilot-training success: A meta-analysis of published research. *The International Journal of Aviation Psychology*, *4*(4), 297-313.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72-98.

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*(1), 28-42.

*Hunter, D. R., & Thompson, N. A. (1978). *Pilot Selection System Development* (No. AFHRL-TR-78-33). Brooks AFB, San Antonio, TX: Human Resources Laboratory.

Jacobs, D., Niemczyk, M., Nullmeyer, R., Cooke, N., & Cline, P. (2020). Depression, anxiety and stress in collegiate aviators. *The Collegiate Aviation Review International, 38*(1), 46-68.

Jak, S., & Cheung, M. W. L. (2019). Meta-analytic structural equation modeling with moderating effects on SEM parameters. *Psychological Methods*.

James, M., & Carretta, T. R. (2002). g2k. *Human Performance*, *15*(1-2), 3-23.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1998). *The g factor: The science of mental ability* (Vol. 648). Westport, CT: Praeger.

Jensen, A. R., & Weng, L. -J. (1994). What is a good g? *Intelligence*, 18,231–258.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35, 1–19.

Johnson, J. F., Barron, L. G., Carretta, T. R., & Rose, M. R. (2017). Predictive validity of spatial ability and perceptual speed tests for aviator training. *The International Journal of Aerospace Psychology*, *27*(3-4), 109-120.

Johnson, W., & Bouchard Jr, T. J. (2005). Constructive replication of the visual–perceptual-image rotation model in Thurstone's (1941) battery of 60 tests of mental ability. *Intelligence, 33*(4), 417-430.

Johnston, P. J., & Catano, V. M. (2013). Investigating the validity of previous flying experience, both actual and simulated, in predicting initial and advanced military pilot training performance. *The International Journal of Aviation Psychology*, *23*(3), 227-244.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Kahana, S. Y., Youngstrom, E. A., & Glutting, J. J. (2002). Factor and subtest discrepancies on the differential ability scales: Examining prevalence and validity in predicting academic achievement. *Assessment*, *9*(1), 82-93.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, *74*(4), 657.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. Katz, L. (2006). Finding the right stuff. *US Army Research Institute*. Fort Rucker, AL: U.S. Army.

Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. *School Psychology Quarterly*, 14, 239–262.

Kell, H. J., & Lang, J. W. (2017). Specific abilities in the workplace: More important than g?. *Journal of Intelligence*, *5*(2), 13.

Kell, H. J., & Lang, J. W. (2018). The great debate: General ability and specific abilities in the prediction of important outcomes. *Journal of Intelligence*, *6*(3), 1-8.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1, 4th ed., pp. 233–265). Boston, MA: McGraw-Hill.

King, R. E., Carretta, T. R., Retzlaff, P., Barto, E., Ree, M. J., & Teachout, M. S. (2013). Standard cognitive psychological tests predict military pilot training outcomes. *Aviation Psychology and Applied Human Factors*, *3*(1), 28–38.

Kirschner, J. (2011). The stress coping skills of undergraduate collegiate aviators (Master Thesis). Retrieved from https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1060&context=techmasters

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kock, F. D., & Schlechter, A. (2009). Fluid intelligence and spatial reasoning as predictors of pilot training performance in the South African Air Force (SAAF). *SA Journal of Industrial Psychology*, *35*(1), 31-38.

Koran, J. (2020). Indicators per Factor in Confirmatory Factor Analysis: More is not Always Better. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-8.

Kubisiak, C. & Katz, L. (2006). *U.S. Army Aviator job analysis* (ARI TR 1189). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*, *19*(6), 339-345.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: implications for graduate student selection and performance. *Psychological Bulletin*, *127*(1), 162-181.

Lang, J. W., & Kell, H. J. (2019). General mental ability and specific abilities: Their relative importance for extrinsic career success. *Journal of Applied Psychology*. Advance online publication.

Lang, J. W., Kersting, M., Hülsheger, U. R., & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: The perspective of the nested-factors model of cognitive abilities. *Personnel Psychology*, *63*(3), 595-640.

Legree, P. J., Pifer, M. E., & Grafton, F. C. (1996). Correlations among cognitive abilities are lower for higher ability groups. *Intelligence*, *23*(1), 45-57.

Light, A., & McGee, A. (2015). Employer learning and the "importance" of skills. *Journal of Human Resources*, *50*(1), 72-107.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

Lochner, K., & Nienhaus, N. (2016). *The predictive power of assessment for pilot selection*. Technical report published by cut-e Group. Retrieved April 13, 2020, from https://www.tts-talent.com/wp-content/uploads/2016/11/images_Blog-Content_cut-e_White_Paper_Pilot_Assessment.pdf

Lynch, W. E. (1991). *A Meta-analysis of pilot selection tests: Success and performance in pilot training* (No. AFIT/GLM/LSM/91S-44). Wright-Patterson AFB, OH: Air Force Institute of Technical, School of Systems and Logistics.

Lyons, B. D., Hoffman, B. J., & Michel, J. W. (2009). Not much more than g? An examination of the impact of intelligence on NFL performance. *Human Performance*, *22*(3), 225-245.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130-149.

Marcoulides, K. M., & Yuan, K. H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 148-153.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*(2), 181–220.

Martinussen, M. (1996). Psychological measures as predictors of pilot performance: A meta-analysis. *The International Journal of Aviation Psychology*, *6*(1), 1-20.

Martinussen, M., & Torjussen, T. (1998). Pilot selection in the Norwegian Air Force: A validation and meta-analysis of the test battery. *The International Journal of Aviation Psychology*, *8*(1), 33-45.

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, *2*(4), 260-293.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of job performance. *Journal of Applied Psychology*, *73*(2), 327-330.

McFarland, D. J. (2016). Modeling general and specific abilities: Evaluation of bifactor models for the WJ-III. *Assessment*, *23*(6), 698-706.

McGrew, K. S., Keith, T. Z., Flanagan, D. P., & Vanderwood, M. (1997). Beyond g: The impact of Gf-Gc specific cognitive ability research on the future use and interpretation of intelligence tests in the schools. *School Psychology Review*, 26, 189–201.

Molenaar, D. (2016). On the distortion of model fit in comparing the bifactor model and the higher-order factor model. *Intelligence*, *57*, 60-63.

Morgan, G. B. (2015). Mixed mode latent class analysis: An examination of fit index performance for classification. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 76-86.

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*, *3*(1), 2-20.

Murphy, K. R. (1989). Is the relationship between cognitive ability and job performance stable over time? *Human Performance*, *2*(3), 183-200.

Murphy, K. (2017). What can we learn from "not much more than g"? *Journal of Intelligence*, *5*(1), 8.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*(5), 407-422.

Muthén, B. (2015). General and specific factors in selection modeling. In M. Rosén, K. Yang Hansen, & U. Wolff (Eds.), *Cognitive abilities and educational outcomes, methodology of educational measurement and assessment* (pp. 223–236). Basel, Switzerland: Springer

Nachtigall, C., Kroehne, U., Funke, F., & Steyer, R. (2003). Pros and cons of structural equation modeling. *Methods Psychological Research Online*, *8*(2), 1-22.

Nye, C. D., Chernyshenko, O. S., Stark, S., Drasgow, F., Phillips, H. L., Phillips, J. B., & Campbell, J. S. (2020). More than g: Evidence for the Incremental Validity of Performance-Based Assessments for Predicting Training Performance. *Applied Psychology*, *69*(2), 302-324.

O'Connor, P., Campbell, J., Newon, J., Melton, J., Salas, E., & Wilson, K. A. (2008). Crew resource management training effectiveness: A meta-analysis and some critical needs. *The International Journal of Aviation Psychology, 18*(4), 353-368.

Oh, H. J., Glutting, J. J., Watkins, M. W., Youngstrom, E. A., & McDermott, P. A. (2004). Correct interpretation of latent versus observed abilities: Implications from structural equation modeling applied to the WISC-III and WIAT linking sample. *The Journal of Special Education*, *38*(3), 159-173.

*Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than g. *Journal of Applied Psychology*, *79*(6), 845-851.

Olson, T. M., Walker, P. B., & Phillips, IV, H. L. (2009). Assessment and selection of aviators in the US military. In P. E. O'Connor & J. V. Cohn (Eds.), *Human performance enhancement in high-risk environments: Insights, developments, and future directions from military research* (pp. 37–57). Santa Barbara, CA: Praeger Security International.

Ones, D. S., Dilchert, S., & Viswesvaran, C. (2012). Cognitive abilities. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 179–224). New York: Oxford University Press.

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, *17*(6), 609-626.

Paullin, C., Katz, L. C., Bruskiewicz, K. T., Houston, J., & Damos, D. (2006). *Review of aviator selection* (No. TR-493). Minneapolis, MN: Personnel Decisions Research Inst.

Paunonen, S. V., & Hong, R. Y. (2010). Self-Efficacy and the Prediction of Domain-Specific Cognitive Abilities. *Journal of Personality*, *78*(1), 339-360.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). New York: Harcourt Brace.

Phillips, J. B., Chernyshenko, O. S., Stark, S., Drasgow, F., & Phillips, I. V. (2011). *Development of scoring procedures for the Performance Based Measurement (PBM) test: Psychometric and criterion validity investigation* (No. NAMRU-D-12-10). Dayton, OH: Naval Medical Research Unit.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93-115.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement*, *27*(1), 52-71.

Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing*, *6*(1), 81-95.

Ree, M. J., Carretta, T. R., & Steindl, J. R. (2001). Cognitive ability. In N. Anderson, D. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology: Personnel psychology*, (Vol. 1, pp. 219–232). London/New York: Sage.

*Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior knowledge in complex training performance. *Journal of Applied Psychology*, *80*(6), 721-730.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than g. *Personnel Psychology*, *44*(2), 321-332.

Ree, M. J., & Earles, J. A. (1991). The stability of g across different methods of estimation. *Intelligence*, *15*, 271–278.

Ree, M.J. & Earles, J.A. (1996) Predicting Occupational Criteria: Not Much More than g. In Dennis, I. and Tapsfield, P. (eds.), *Human Abilities: Their Nature and Measurement*. Erlbaum, New Jersey: NJ.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology*, *79*(4), 518-524.

Reeve, C. L., & Blacksmith, N. (2009). Identifying g: A review of current factor analytic practices in the science of mental abilities. *Intelligence*, 37,487–494.

Reeve, C. L., & Bonaccio, S. (2011). Nature and structure of intelligence. In T. Chamorro-Premuzic, A. Furnham, & S. von Stumm (Eds.), *Handbook of individual differences* (pp. 187–216). Oxford, UK: Wiley-Blackwell.

Reeve, C. L., Scherbaum, C., & Goldstein, H. (2015). Manifestations of intelligence: Expanding the measurement space to reconsider specific cognitive abilities. *Human Resource Management Review*, *25*(1), 28-37.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*(6), 544-559.

Rhone, J. M. (2008). *Battle Management as a Basic Air Force Doctrine Operational Function* (Master Thesis). Fort Leavenworth, KS: Army Command and General Staff College.

Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer training school selection decisions. *Military Psychology*, *8*(2), 95-113.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: calculating and interpreting statistical indices. *Psychological methods*, *21*(2), 137-150.

Rosenthal, R., & DiMatteo, M. R. (2000). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: an expanded typology. *Journal of Applied Psychology*, *85*(1), 112-118.

Salgado, J.F. (2017). Bandwidth-fidelity dilemma BT. In V. Zeigler-Hill, & T.K. Shackelford (Eds.), *Encyclopedia of personality and individual differences*. Cham: Springer.

Sass, D. A., & Smith, P. L. (2006). The effects of parceling unidimensional scales on structural parameter estimates in structural equation modeling. *Structural Equation Modeling*, *13*(4), 566-586.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance*, *15*(1-2), 187-210.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262-274.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162-173.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*(3), 432-439.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53-61.

Schneider, D., & Dorans, N. (1999). Concordance between SAT I and ACT scores for individual students. *Research Notes (RN-07)*. New York: The College Board.

Schneider, W. J., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review*, *25*(1), 12-27.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. psychology press.

Schwarz, G. (1978). Estimating the dimensions of a model. Annals of Statistics, 6, 461– 464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.

Shappell, S. A. & Wiegmann, D. A. (1996). U.S. Naval aviation mishaps, 1977-92: Differences between single- and dual- piloted aircraft. *Aviation, Space, and Environmental Medicine*, 67, 65-69.

Shepard, L. A. (1993). Chapter 9: Evaluating test validity. *Review of Research in Education*, *19*(1), 405-450.

Shi, L. (2014). *Validity and Reliability of a Mini-Situational Judgment Test (SJT) for Pilots* (Master Thesis). Retrieved from https://repository.lib.fit.edu/handle/11141/275

Silvia, P. J., Thomas, K. S., Nusbaum, E. C., Beaty, R. E., & Hodges, D. A. (2016). How does music training predict cognitive abilities? A bifactor approach to musical expertise and intelligence. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(2), 184-190.

Sidik, K., & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, *26*(9), 1964-1981.

Skinner, J., & Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and Factor Analysis of Form O* (No. AFHRL-TR-86-68). Brooks AFB, San Antonio, TX: Air Force Human Resources Laboratory.

Sloan, T. A., Lundin, M., Wilson, D., & Robinnette, R. (2018). The use of test anxiety assessment and anxiety reduction training to predict and improve performance of collegiate pilot trainees. *The Collegiate Aviation Review International*, *28*(2), 60-68.

Smith, G., Bjerke, E., Smith, M., Christensen, C., Carney, T., Craig, P., & Niemczyk, M. (2016). Pilot source study 2015: An analysis of FAR Part 121 pilots hired after Public Law 111-216—Their backgrounds and subsequent successes in US regional airline training and operating experience. *Journal of Aviation Technology and Engineering*, *6*(1), 64-89.

Smith, G. M., Herchko, D., Bjerke, E., Niemczyk, M., Nullmeyer, R., Paasch, J., & NewMyer, D. A. (2013). The 2012 pilot source study (Phase III): Response to the pilot certification and qualification requirements for air carrier operations. *Journal of Aviation Technology and Engineering*, *2*(2), 13-23.

Smith, G. M., NewMyer, D. A., Bjerke, E., Niemczyk, M., & Hamilton, R. A. (2010). Pilot source study: An analysis of pilot backgrounds and subsequent success in US regional airline training programs. *International Journal of Applied Aviation Studies*, *10*(1), 73-96.

Spearman, C. E. (1904). 'General intelligence' objectively determined and measured. *American Journal of Psychology, 5*, 201-293.

Stanek, K. C., & Ones, D. S. (2018). Taxonomies and compendia of cognitive ability and personality constructs and measures relevant to industrial, work and organizational psychology. *The SAGE handbook of industrial, work and organizational psychology (2nd ed., Vol. 1, pp. 366-407). Thousand Oaks, CA: SAGE*, *10*(9781473914940), n14.

Stankov, L. (2017). Overemphasized "g". *Journal of Intelligence*, *5*(4), 33.

Stokes, A. & Kite, K. (1994). *Flight stress: Stress, fatigue, and performance*. Cambridge: University Press.

Street, D. R., Helton, K. T., and Dolgin D. L. (1992). *The unique contribution of selected personality tests to the prediction of success in naval pilot training* (NAMRL- 1374). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Stricker, L. J. (2005). The biographical inventory in naval aviation selection: Inside the black box. *Military Psychology*, *17*(1), 55-67.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197–208.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior*, 29, 332–339.

Thorndike, R. L. (1987). Stability of factor loadings. *Personality and Individual Differences, 8*, 585–586.

Thurstone, L. L. (1938). *Primary mental abilities* (Vol. 119). Chicago: University of Chicago Press.

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology*, *26*(1), 1-9.

Tucker, L. R, & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.

Ullman, J., & Bentler, P. (2012). Structural equation modeling. *Handbook of Psychology*, (2nd ed., pp. 661 – 683). New Jersey: John Wiley & Sons, Inc.

Vaden, E. A., & Hall, S. (2005). The effect of simulator platform motion on pilot training transfer: A meta-analysis. *The International Journal of Aviation Psychology*, 15(4), 375-393.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., ... & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods, 7*(1), 55-79.

Valerius, S., & Sparfeldt, J. R. (2014). Consistent g-as well as consistent verbal-, numerical-and figural-factors in nested factor models? Confirmatory factor analyses using three test batteries. *Intelligence*, *44*, 120-133.

Van Buskirk, S. L. (2018). *Triangular love: 'not much more than G'* (Order No. 13420496). Available from ProQuest Dissertations & Theses Global. (2158353973). Retrieved from https://search.proquest.com/docview/2158353973?accountid=13360

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321-327.

Vernon, P. E. (1961). *The structure of human abilities* (2nd ed.). London, UK: Methuen

Viswesvaran, C., & Ones, D. S. (2002). Agreements and disagreements on the role of general mental ability (GMA) in industrial, work, and organizational psychology. *Human Performance*, *15*(1-2), 211-231.

Walters, L. C., Miller, M. R., & Ree, M. J. (1993). Structured interviews for pilot selection: No incremental validity. *The International Journal of Aviation Psychology*, *3*(1), 25-38.

Wang, H., Su, Y., Shang, S., Pei, M., Wang, X., & Jin, F. (2018). Working memory: A criterion of potential practicality for pilot candidate selection. *The International Journal of Aerospace Psychology*, *28*(3-4), 64-75.

Wee, S. (2018). Aligning predictor-criterion bandwidths: Specific abilities as predictors of specific performance. *Journal of Intelligence*, *6*(3), 53-66.

Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than g: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology*, *99*(4), 547-563.

Wheeler, J. L., & Ree, M. J. (1997). The role of general and specific psychomotor tracking ability in validity. *International Journal of Selection and Assessment*, *5*(2), 128-136.

Williams, H. P., Albert, A. O., & Blower, D. J. (2000). *Selection of officers for US naval aviation training*. Pensacola, FL: Naval Aerospace Medical Research Lab.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under-and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, *1*(4), 354-365.

Woychesin, D. E. (2002). Validation of the Canadian Automated Pilot Selection System (CAPSS) against primary flying training results. *Canadian Journal of Behavioural Science*, *34*(2), 84-91.

Yacavone, D. W. (1993). Mishap trends and cause factors in naval attrition: A review of naval safety center data, 1986-90. *Aviation, Space, and Environmental Medicine*, *64*, 392-395.

Yazgan, E., Çilingir, F. C., Erol, D., & ANAGÜN, A. S. (2017). An Analysis of the Factors Influencing Score Achieved during Pilot Training. *Transactions Of The Japan Society For Aeronautical And Space Sciences*, *60*(4), 202-211.

Yoon, M., & Lai, M. H. (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 201-213.

Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2020). Using bifactor models to examine the predictive validity of hierarchical constructs: Pros, cons, and solutions. *Organizational Research Methods*. https://doi.org/10.1177/1094428120915522

Ziegler, M., & Peikert, A. (2018). How specific abilities might throw 'g'a curve: An idea on how to capitalize on the predictive validity of specific cognitive abilities. *Journal of Intelligence*, *6*(3), 32-52.

Zierke, O. (2014). Predictive validity of knowledge tests for pilot training outcome. *Aviation Psychology and Applied Human Factors, 2*(4), 98-105.

APPENDIX A

The Air Force Officer Qualifying Test (AFOQT) Test Items

(The content of this appendix is extracted with permission from *Military Flight Aptitude Tests*, by Solomon Wiener, 2005, Lawrenceville, NJ: Thomson Paterson's. Copyright 2005 by Thomson Learning™)

This appendix includes further information about the AFOQT and examples for test items typically found in the AFOQT subtests. Despite the continued revision of the AFOQT, the developers persevere to keep the measured constructs underlying its subtests comparable, with an extensive equating effort when introducing newer versions of the test.

Table A-1.
Format of the Specimen AFOQT

Subtest		**No of Items**
Verbal Analogies***	VA	25
Reading Comprehension***	RC	25
Word Knowledge***	WK	25
Arithmetic Reasoning***	AR	25
Data Interpretation*	DI	25
Math Knowledge***	MK	25
Mechanical Comprehension*	MC	20
Electrical Maze*	EM	20
Block Counting***	BC	30
Rotated Blocks*	RB	15
Hidden Figures*	HF	15
Scale Reading*	SR	40
Table Reading***	TR	40
Instrument Comprehension***	IC	25
Aviation Information***	AI	20
General/Physical Science***	GS/PS	20

Note. subtests indicated by (***) are those currently operational in the present Form T; subtests indicated by (*) are currently not used.

VERBAL ANALOGIES

Directions

This part of the test has 25 questions designed to measure your ability to reason and see relationships between words. Each question begins with a pair of capitalized words. You are to choose the choice that best completes the analogy developed at the beginning of each question. That is, select the choice that shows a relationship similar to the one shown by the original pair of capitalized words. Then, mark the space on your answer form that has the same number and letter as your choice.

Now look at the two sample questions below.

1. FINGER is to HAND as TOOTH is to

(A) TONGUE
(B) LIPS
(C) NOSE
(D) MOUTH
(E) MOLAR

The correct answer is (D). A *finger* is part of the *hand*; a *tooth* is part of the *mouth*.

2. RACQUET is to COURT as

(A) TRACTOR is to FIELD
(B) BLOSSOM is to BLOOM
(C) STALK is to PREY
(D) PLAN is to STRATEGY
(E) MOON is to PLANET

The correct answer is (A). A *racquet* is used (by a tennis player) on the *court*; a *tractor* is used (by a farmer) on the *field*.

ARITHMETIC REASONING

Directions

This part of the test has 25 questions that measure mathematical reasoning or your ability to arrive at solutions to problems. Each problem is followed by five possible answers. Decide which one of the five choices is most nearly correct. Then, mark the space on your answer form that has the same number and letter as your choice. Use the scratch paper that has been given to you to do any figuring.

Now look at the two sample problems below.

1. A field with an area of 420 square yards is twice as large in area as a second field. If the second field is 15 yards long, how wide is it?

(A) 7 yards
(B) 14 yards
(C) 28 yards
(D) 56 yards
(E) 90 yards

The correct answer is (B). The second field has an area of 210 square yards. If one side is 15 yards, the other side must be 14 yards ($15 \times 14 = 210$).

2. An applicant took three typing tests. The average typing speed on these three tests was 48 words per minute. If the applicant's speed on two of these tests was 52 words per minute, what was the applicant's speed on the third test?

(A) 46 words per minute
(B) 44 words per minute
(C) 42 words per minute
(D) 40 words per minute
(E) 38 words per minute

The correct answer is (D). The formula for finding an average is as follows:

average = sum of terms/numbers of terms

In this case the problem provides the average (48), two of the terms (52 + 52), and the number of terms (3). Substitute this information into the formula for average and then solve for *x* (the missing term).

$$48 = \frac{52 + 52 + x}{3}$$
$$48 \times 3 = 104 + x$$
$$144 = 104 + x$$
$$40 = x$$

READING COMPREHENSION

Directions

This part of the test has 25 questions designed to measure your ability to read and understand paragraphs. For each question, you are to select the answer that best completes the statement or answers the question based on the information contained in the passage. Then, mark the space on your answer form that has the same number and letter as your choice.

Here are two sample questions:

1. Because of our short life span of seventy-odd years, it is easy for human beings to think of Earth as a planet that never changes. Yet we live on a dynamic planet with many factors contributing to change. We know that wind and rain erode and shape our planet. Many other forces are also at work, such as volcanic activity, temperature fluctuations, and even extraterrestrial interaction such as meteors and gravitational forces. The earth, in actuality, is a large rock

(A) in a state of inertia.
(B) that is quickly eroding.
(C) that is evolving.
(D) that is subject to temperature fluctuations caused by interplanetary interaction.
(E) that is subject to winds caused by meteor activity.

The correct answer is (C). Of the choices given, only choice (C) can be implied from the passage.

2. One theory that explains the similarities between Mayan art and ancient Chinese art is called "diffusion." This theory evolves from the belief that invention is so unique that it happens only once, then is "diffused" to other cultures through travel, trade, and war. This theory might explain why

(A) the airplane and birds both have wings.
(B) certain artifacts in Central America resemble those found in Southeast Asia.
(C) most great art comes from Europe, where there is much travel between countries.
(D) rivers in South America and Africa have some similar features.
(E) England, being so remote in the Middle Ages, is the only country to have castles.

The correct answer is (B). Of the choices given, choice (B) is the only one that the theory might explain.

DATA INTERPRETATION

Directions

This part of the test has 25 questions designed to measure your ability to interpret data from tables and graphs. Each question is followed by four or five possible answers. Decide which answer is correct, then mark the space on your answer form that has the same number and letter as your choice.

Study the two sample questions below before you begin the Data Interpretation Test.

Your score on this test is based on the number of questions you answer correctly. You should try to answer every question. You will not lose points or be penalized for guessing. Do not spend too much time on any one question.

When you begin, be sure to start with question number 1 of Part 4 of your test booklet and number 1 of Part 4 on your answer sheet.

Number of days absent per employee (sickness)	1	2	3	4	5	6	7	8 or more
Number of employees	76	23	6	3	1	0	1	0

Total Number of Employees: 400
Period Covered: January 1, 2001—December 31, 2001

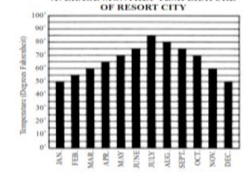
1. Based on the data shown above, the total number of working days lost due to sickness in 2001 was

(A) 110
(B) 137
(C) 144
(D) 158
(E) 164

The correct answer is (E). Multiplying the number of employees by the number of days absent per employee (sickness) and then adding the products, we arrive at

$$76 + 46 + 18 + 12 + 5 + 0 + 7 + 0 = 164.$$

AVERAGE MONTHLY TEMPERATURE OF RESORT CITY



2. Based on the information in the graph above, the average monthly temperature in November is the same as in

(A) January.
(B) February.
(C) March.
(D) April.
(E) May.

The correct answer is (C). The average temperature in November is 60°F. The only other month in which the average temperature is 60°F is March.

WORD KNOWLEDGE

Directions

This part of the test has 25 questions designed to measure verbal comprehension involving your ability to understand written language. For each question, you are to select the option that means the same or most nearly the same as the capitalized word. Then mark the space on your answer form that has the same number and letter as your choice.

Here are two sample questions:

1. CRIMSON:

(A) bluish
(B) colorful
(C) crisp
(D) lively
(E) reddish

The correct answer is (E). *Crimson* means "a deep purple red." Choice (E) has almost the same meaning. None of the other options has the same or a similar meaning.

2. CEASE:

(A) continue
(B) fold
(C) start
(D) stop
(E) transform

The correct answer is (D). *Cease* means "to stop." Choice (D) is the only option with the same meaning.

MATH KNOWLEDGE

Directions

This part of the test has 25 questions designed to measure your ability to use learned mathematical relationships. Each problem is followed by five possible answers. Decide which one of the five choices is most nearly correct. Then, mark the space on your answer form that has the same number and letter as your choice. Use scratch paper to do any figuring.

Here are three sample questions:

1. The reciprocal of 5 is

(A) 0.1
(B) 0.2
(C) 0.5
(D) 1.0
(E) 2.0

The correct answer is (B). The reciprocal of 5 is $\frac{1}{5}$ or 0.2.

2. The expression "3 factorial" equals

(A) $\frac{1}{9}$
(B) $\frac{1}{6}$
(C) 6
(D) 9
(E) 27

The correct answer is (C). "3 factorial" or 3! equals $3 \times 2 \times 1 = 6$.

3. The logarithm to the base 10 of 1,000 is

(A) 1
(B) 1.6
(C) 2
(D) 2.7
(E) 3

The correct answer is (E). $10 \times 10 \times 10 = 1{,}000$. The logarithm of 1,000 is the exponent 3 to which the base 10 must be raised.

MECHANICAL COMPREHENSION

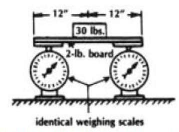
Directions

This part of the test has 20 questions designed to measure your ability to learn and reason with mechanical terms. Included in this part of the test are diagrams of mechanical devices. Preceding the diagrams are questions or incomplete statements. Study each diagram carefully and select the choice that best answers the question or completes the statement. Then, mark the space on your answer form that has the same number and letter as your choice.

Here are two sample questions:

1. In the figure shown, the weight held by the board and placed on the two identical scales will cause *each* scale to read

(A) 8 lbs.
(B) 15 lbs.
(C) 16 lbs.
(D) 30 lbs.
(E) 32 lbs.



The correct answer is (C). 30 lbs. + 2 lbs. = 32 lbs., the total weight equally supported by two scales. $\frac{32}{2}$ = 16 lbs., the reading on each scale.

2. In the figure shown, the pulley system consists of a fixed block and a movable block. The theoretical mechanical advantage is

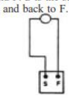
(A) 1
(B) 2
(C) 3
(D) 4
(E) 5

The correct answer is (B). The number of parts of the rope going to and from the movable block indicates the mechanical advantage. In this case it is 2.

ELECTRICAL MAZE

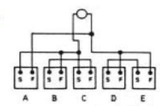
Directions

This part of the test has 20 questions designed to measure your ability to choose a correct path from among several choices. In the picture below is a box with dots marked S and F. S is the starting point, and F is the finishing point. Follow the line from S, through the circle at the top of the picture, and back to F.



In the problems in this test, there will be five such boxes. Only *one* box will have a line from the S, through the circle, and back to the F in the same box. Dots on the lines show the *only* places where turns or direction changes can be made between lines. If lines meet or cross where there is *no dot*, turns or direction changes *cannot be made*. Now try sample problem 1:

1.

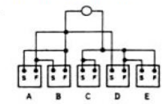


The correct answer is (A). The first box is the one that has the line from S, through the circle, and back to F.

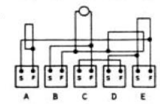
Each diagram in the test has only one box that has a line through the circle and back to F. Some lines are wrong because they lead to a dead end. Some lines are wrong because they come back to the box without going through the circle. Some lines are wrong because they lead to other boxes. Some are wrong because they retrace the same line.

Now try sample problems 2 and 3:

2.



3.



For sample problem 2, the correct answer is (D).

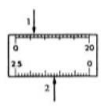
For sample problem 3, the correct answer is (B).

SCALE READING

Directions

This part of the test has 40 questions designed to measure your ability to read scales, dials, and meters. You are given a variety of scales on which various points are indicated by numbered arrows. Estimate the numerical value indicated by each arrow, find the choice closest to this value in the item having the same numbers as the arrow, and then mark your answer on the answer sheet. Now look at the sample items below:

1. (A) 6.00
(B) 5.00
(C) 4.25
(D) 2.25
(E) 1.25

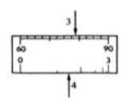


2. (A) 13.0
(B) 12.0
(C) 10.2
(D) 1.3
(E) 1.2

For sample item 1, the correct answer is (B). There are five subdivisions of four steps each between 0 and 20. The arrow points between the long subdivision markers representing 4 and 8. Since it points to the marker that is one step to the right of subdivision marker 4, it points to 5.00.

For sample item 2, the correct answer is (E). The scale runs from right to left. There are five subdivisions of five steps each, so each step represents .1, and the arrow points to the marker representing 1.2.

3. (A) 81.75
(B) 79.5
(C) 78.75
(D) 77.60
(E) 67.50



4. (A) 1.75
(B) 1.65
(C) 1.50
(D) .75
(E) .65

For sample item 3, the correct answer is (C). The arrow points between two markers. You must estimate the fractional part of the step as accurately as possible. Since the arrow points halfway between the markers representing 77.5 and 80.0, it points to 78.75.

For sample item 4, the correct answer is (E). Each step represents .5, but the steps are of unequal width, with each step being two-thirds as wide as the preceding one. Therefore, the scale is compressed as the values increase. The arrow is pointing to a position halfway between the marker representing .5 and 1.0, but because of the compression of the scale the value of this point must be less than .75. Actually, it is .65, which is choice (E).

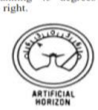
INSTRUMENT COMPREHENSION

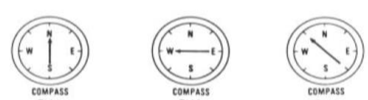
Directions

This part of the test has 20 questions designed to measure your ability to determine the position of an airplane in flight from reading instruments showing its compass heading, its amount of climb or dive, and its degree of bank to right or left. In each item, the left-hand dial is labeled ARTIFICIAL HORIZON. On the face of this dial the small aircraft silhouette remains stationary, while the positions of the heavy black line and the black pointer vary with changes in the position of the airplane in which the instrument is located.

How to Read the Artificial Horizon Dial

The heavy black line represents the HORIZON LINE. The black pointer shows the degree of BANK the to right or left. The HORIZON LINE tilts as the aircraft is banked and is always at a right angle to the pointer.

Dial 1 shows an airplane neither climbing nor diving, with no bank.	Dial 2 shows an airplane climbing and banking 45 degrees to the pilot's right.	Dial 3 shows an airplane diving and banked 45 degrees to the pilot's left.
		
Dial 1	Dial 2	Dial 3
If the airplane is neither climbing nor diving, the horizon line is directly on the silhouette's fuselage. If the airplane has no bank, the black pointer is seen to point to zero.	If the airplane is climbing, the fuselage silhouette is seen between the horizon line and the pointer. The greater the amount of climb, the greater the distance between the horizon line and the fuselage silhouette. If the airplane is banked to the pilot's right, the pointer is seen to the left of zero.	If the airplane is diving, the horizon line is seen between the fuselage silhouette and the pointer. The greater the amount of dive, the greater the distance between the horizon line and the fuselage silhouette. If the airplane is banked to the pilot's left, the pointer is seen to the right of zero.



On each item, the right-hand dial is labeled COMPASS. On this dial, the arrow shows the compass direction in which the airplane is headed at the moment. Dial 4 shows it headed north; dial 5 shows it headed west; and dial 6 shows it headed northwest.

A Sample Question Explained

Each item in this test consists of two dials and four silhouettes of airplanes in flight. Your task is to determine which one of the four airplanes is MOST NEARLY in the position indicated by the two dials. YOU ARE ALWAYS LOOKING NORTH AT THE SAME ALTITUDE AS EACH OF THE PLANES. EAST IS ALWAYS TO YOUR RIGHT AS YOU LOOK AT THE PAGE. Item X is a sample. In item X the dial labeled ARTIFICIAL HORIZON shows that the airplane is NOT banked, and is neither climbing nor diving. The COMPASS shows that it is headed southeast. The only one of the four airplane silhouettes that meets these specifications is in the box lettered (C), so the answer to X is (C). Note that (B) is a rear view, while (D) is a front view. Note also that (A) is banked to the right and that (B) is banked to the left.

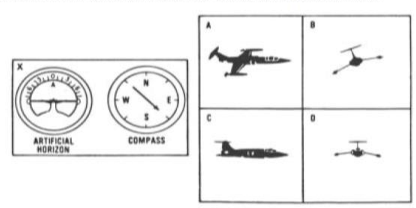


TABLE READING

Directions

This part of the test has 40 questions designed to test your ability to read tables quickly and accurately.

Now, look at the following sample items based on the tabulation of turnstile readings shown below:

TABULATION OF TURNSTILE READINGS

Turnstile Number	Turnstile Readings At 5:30 a.m.	6:00 a.m.	7:00 a.m.	8:00 a.m.	9:00 a.m.	9:30 a.m.
1	79078	79090	79225	79590	79860	79914
2	24915	24930	25010	25441	25996	26055
3	39509	39530	39736	40533	41448	41515
4	58270	58291	58396	58958	59729	59807
5	43371	43378	43516	43888	44151	44217

For each question, determine the turnstile reading for the turnstile number and time given. Choose as your answer the letter of the column in which the correct reading is found.

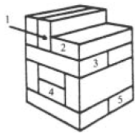
	Turnstile Number	Time	A	B	C	D	E
1.	1	8:00 a.m.	25441	25996	79225	79590	79860
2.	2	6:00 a.m.	24915	24930	25010	39530	79090
3.	4	9:30 a.m.	41515	44217	44151	59729	59807
4.	5	7:00 a.m.	39530	39736	43516	58291	58396
5.	3	5:30 a.m.	39509	39530	39736	58270	58291

The correct answers are 1. (D); 2. (B); 3. (E); 4. (C); and 5. (A).

BLOCK COUNTING

Directions

This part of the test has 20 questions designed to measure your ability to "see into" a three-dimensional pile of blocks and determine how many pieces are touched by certain numbered blocks. *All of the blocks in each pile are the same size and shape.* A block is considered to touch the numbered block if any part, even a corner or an edge, touches. Look at the sample below:



	KEY				
Block	**A**	**B**	**C**	**D**	**E**
1	1	2	3	4	5
2	3	4	5	6	7
3	5	6	7	8	9
4	2	3	4	5	6
5	2	3	4	5	6

For sample problem 1, the correct answer is (D). Block 1 touches the other 2 top blocks and the 2 blocks directly below it. The total number of blocks touched by 1 is, therefore, 4.

For sample problem 2, the correct answer is (A). Block 2 touches blocks 1 and 3, and the unnumbered block to the right of block 3. Since block 2 touches 3 other blocks, the answer is 3.

For sample problem 3, the correct answer is (C). Now look at sample problem 3. It touches 3 blocks above, 3 blocks below, and one block on the right. Therefore, the correct answer is 7.

For sample problem 4, the correct answer is (D). Now count the blocks touching blocks 4 and 5. For block 4, the correct answer is 5. For sample problem 5, the correct answer is (C).

Your score on this test is based on the number of questions you answer correctly. You should try to answer every question. You will not lose points or be penalized for guessing. Do not spend too much time on any one question.

When you begin, be sure to start with question number 1 of Part 11 of your test booklet and number 1 of Part 11 on your answer sheet.

ROTATED BLOCKS

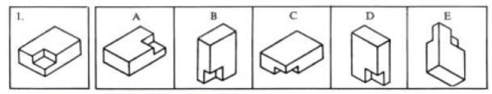
Directions

This part of the test has 15 questions designed to measure your ability to visualize and manipulate objects in space. In each item, you are shown a picture of a block. You must find a second block that is just like the first.

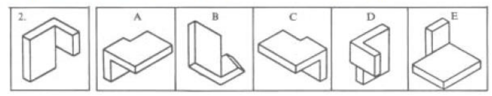
Look at the two blocks below. Although viewed from different points, the blocks are the same.

Look at the two blocks below. They are not alike. They can never be turned so that they will be alike.

Now look at the sample item below. Which of the five choices is just like the first block?



For sample problem 1, the correct answer is (D). It is the same block as seen from a different side.



For sample problem 2, the correct answer is (C).

AVIATION INFORMATION

Directions

This part of the test has 20 questions designed to measure your knowledge of aviation. Each of the questions or incomplete statements is followed by five choices. Decide which one of the choices best answers the question or completes the statement.

Now look at the two sample questions below:

1. The force necessary to overcome gravitational force to keep the airplane flying is termed

(A) power.
(B) drag.
(C) lift.
(D) thrust.
(E) weight.

The correct answer is (C). To keep the airplane flying, *lift* must overcome the weight or gravitational force.

2. The ailerons are used primarily to

(A) bank the airplane.
(B) control the direction of yaw.
(C) permit a slower landing speed.
(D) permit a steep angle of descent.
(E) control the pitch attitude.

The correct answer is (A). The ailerons, located on the trailing edge of each wing near the outer tip, are used primarily to bank (roll) the airplane around its longitudinal axis. The banking of the wing results in the airplane turning in the direction of the bank.

GENERAL SCIENCE

Directions

This part of the test has 20 questions designed to measure your scientific knowledge. Each of the questions or incomplete statements is followed by five choices. Decide which one of the choices best answers the question or completes the statement.

Now look at the three sample questions below:

1. An eclipse of the sun throws the shadow of the

(A) moon on the sun.
(B) earth on the sun.
(C) sun on the earth.
(D) earth on the moon.
(E) moon on the earth.

2. Substances that hasten a chemical reaction without themselves undergoing change are called

(A) buffers.
(B) catalysts.
(C) colloids.
(D) reducers.
(E) polymers.

3. Lack of iodine is often related to which of the following diseases?

(A) Beriberi
(B) Scurvy
(C) Rickets
(D) Goiter
(E) Asthma

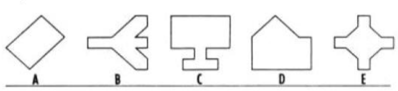
The correct answers are 1. (E); 2. (B); and 3. (D).

HIDDEN FIGURES

Directions

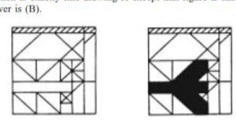
This part of the test has 15 questions designed to measure your ability to see a simple figure in a complex drawing. At the top of each page are five figures, lettered A, B, C, D, and E. Below these on each page are several numbered drawings. You are to determine which lettered figure is contained in each of the numbered drawings.

The lettered figures are:



As an example, look at drawing X below. Which one of the five figures is contained in drawing X?

Now look at drawing Y, which is exactly like drawing X except that figure B has been blackened to show where to look for it. Thus, the correct answer is (B).



Each numbered drawing contains only *one* of the lettered figures. The correct figure in each drawing will always be of the same size and in the same position as it appears at the top of the page. Therefore, do not rotate the page in order to find it. Look at each numbered drawing and decide which one of the five lettered figures is contained in it.

APPENDIX B

Extended Results for the Predictive Studies

This appendix includes extended results to that reported in the result chapter (Chapter 4) for the three predictive studies. To allow for a direct comparison between the results across studies, the presentation was grouped according to the type of statistics reported, rather than the sequence of study. Four tables are presented here including (1) fit statistics of the tested models, (2) factor intercorrelations resulted from the CFA correlated-factor models, (3) factor loadings resulted from the CFA correlated-factor models, (4) subtests loadings on the general and specific factors resulted from the CFA bifactor models.

Table A-1
Summary for the Fit Statistics of the Tested Models

		Model	$\chi 2$	df	CFI	RMSEA	SRMR
Primary		CFA Correlated-factor	1995.63	57	.92	.07	.05
		CFA Bifactor	2039.61	54	.92	.07	.05
		Combined Correlated-factor	2761.44	110	.92	.06	.04
		SEM Bifactor (primary)	2246.43	73	.92	.06	.05
		SEM Bifactor (advanced)	2104.83	73	.93	.06	.04
		SEM Bifactor (academic)	2109.58	61	.92	.07	.05
Cross-validation	Sample 1	CFA Correlated-factor	307.97	47	.91	.07	.05
		CFA Bifactor	312.96	45	.91	.07	.05
		Combined Correlated-factor	328.69	69	.93	.06	.04
		SEM Bifactor (primary)	321.07	51	.91	.07	.05
		SEM Bifactor (advanced)	318.68	51	.91	.07	.05
		SEM Bifactor (Rank)	322.48	51	.91	.07	.05
Cross-validation	Sample 2	CFA Correlated-factor	392.74	47	.94	.06	.06
		CFA Bifactor	357.71	44	.95	.06	.04
		Combined Correlated-factor	365.11	50	.95	.03	.04
		SEM Bifactor (primary)	394.27	51	.95	.06	.04
		SEM Bifactor (advanced)	6168.62	78	.95	.06	.04
		SEM Bifactor (Composite)	404.23	50	.94	.06	.04
Cross-validation	Sample 3	CFA Correlated-factor	521.53	18	.95	.09	.05
		CFA Bifactor	275.23	13	.97	.08	.04
		Combined Correlated-factor	1203.10	261	.94	.06	.03
		SEM Bifactor (primary)	309.56	33	.97	.05	.03
		SEM Bifactor (advanced)	322.19	33	.97	.05	.03
		SEM Bifactor (academic)	730.32	141	.96	.04	.03
Cross-occupation	Pilot	CFA Correlated-factor	397.29	30	.91	.08	.07

		CFA Bifactor	333.25	30	.92	.07	.05
		Combined Correlated-factor	558.47	55	.93	.07	.06
		SEM Bifactor	494.52	54	.94	.07	.05
Cross-occupation	Navigator	CFA Correlated-factor	158.22	30	.93	.07	.06
		CFA Bifactor	135.27	30	.95	.06	.05
		Combined Correlated-factor	186.63	55	.94	.05	.05
		SEM Bifactor	162.03	54	.95	.05	.04
Cross-occupation	ABM	CFA Correlated-factor	138.26	30	.94	.07	.07
		CFA Bifactor	121.74	30	.95	.07	.05
		Combined Correlated-factor	158.68	35	.94	.07	.06
		SEM Bifactor	141.22	34	.95	.07	.05

Table A-2
Summary for the Factors Intercorrelations Resulted from CFA Correlated-factor Models

		Verbal	Quantitative	Spatial	Perceptual	Knowledge
Primary validation	Verbal	1				
	Quantitative	.67	1			
	Spatial	.38	.57	1		
	Perceptual	.39	.80	.61	1	
	Knowledge	.29	.20	.47	.33	1
Cross-validation		Verbal	Quantitative	Spatial	Perceptual	Knowledge
Sample 1	Verbal	1				
	Quantitative	.65	1			
	Spatial	.38	.47	1		
	Perceptual	.32	.58	.66	1	
	Knowledge	.32	.42	.42	.29	1
Cross-validation		Verbal	Quantitative	Spatial	Perceptual	Knowledge
Sample 2	Verbal	1				
	Quantitative	.64	1			
	Spatial	.33	.58	1		
	Perceptual	.37	.87	.68	1	
	Knowledge	.21	.33	.48	.37	1
Cross-validation		Verbal	Quantitative	Spatial	Perceptual	Knowledge
Sample 3	Verbal	1				
	Quantitative	.69	1			
	Spatial	-	-	-		
	Perceptual	-	-	-	-	
	Knowledge	.48	.42	-	-	1
Cross-occupation		Verbal	Quantitative	Spatial	Perceptual	Knowledge
Pilots	Verbal	1				
	Quantitative	.55	1			
	Spatial	.42	.59	1		
	Perceptual	.28	.58	.74	1	
	Knowledge	.16	.19	.45	.29	1
Cross-occupation		Verbal	Quantitative	Spatial	Perceptual	Knowledge
Navigators	Verbal	1				
	Quantitative	.50	1			
	Spatial	.46	.54	1		
	Perceptual	.34	.47	.76	1	
	Knowledge	.29	.26	.55	.36	1
Cross-occupation		Verbal	Quantitative	Spatial	Perceptual	Knowledge
ABM	Verbal	1				
	Quantitative	.52	1			

Spatial	.46	.55	1		
Perceptual	.31	.50	.67	1	
Knowledge	.32	.27	.67	.42	1

Table A-3

Summary for Factor Loadings Based on CFA Correlated-factor Models

Study	Sample	Verbal			Quantitative			Spatial			Perceptual		Knowledge	
		VA	RC	WK	AR	DI	MK	RB	EM	HF	TR	SR/BC	IC	AI
Primary validation	Pilot 0	.79	.71	.68	.78	.67	.69	.60	.49	.48	.55	.67	.69	.59
Cross-validation	Pilot 1	.71	.84	.79	.69	.73	-	.59	.47	.44	.60	.63	.63	.61
Cross-validation	Pilot 2	.73	.85	.80	.73	.75	-	.59	.51	.48	.58	.71	.62	.64
Cross-validation	Pilot 3	.73	.81	.83	.85	.75	.64	-	-	-	-	-	.66	.69
Cross-occupation	Pilot 4	.94	-	.60	.82	-	.69	.51	-	.56	.48	.75*	.62	.64
	Pilot *M*	**.78**	**.80**	**.74**	**.77**	**.73**	**.67**	**.57**	**.49**	**.49**	**.55**	**.69**	**.64**	**.63**
Cross-occupation	Navigator	.91	-	.60	.86	-	.65	.51	-	.54	.48	.79*	.67	.64
Cross-occupation	ABM	.91	-	.70	.81	-	.76	.69	-	.65	.51	.87*	.67	.69

Note. All loadings were significant at $p < .001$; loadings with (*) sign are those from the Block Counting subtest.

Table A-4

Summary for Factor Loadings Based on CFA Bifactor Models

Factor	Study	Sample	Verbal			Quantitative			Spatial			Perceptual		Knowledge	
			VA	RC	WK	AR	DI	MK	RB	EM	HF	TR	SR/BC	IC	AI
Specific Factors	Primary	Pilot 0	.66	.53	.47	-.10	.14	.64	.51	.37	.35	.36	.40	.64	.58
	Cross-V	Pilot 1	.48	.63	.71	.35	.35	-	.45	.38	.30	.44	.44	.55	.55
	Cross-V	Pilot 2	.55	.67	.74	.06	.28	-	.48	.35	.34	.31	.36	.57	.63
	Cross-V	Pilot 3	.30	.50	.60	.77	.17	.25	-	-	-	-	-	.58	.57
	Cross-O	Pilot 4	.80	-	.50	.51	-	.40	.27	-	.30	.31	.49*	.59	.65
		Pilot *M*	**.56**	**.58**	**.60**	**.32**	**.24**	**.43**	**.43**	**.37**	**.32**	**.36**	**.42**	**.59**	**.60**
	Cross-O	Nav	.78	-	.50	.65	-	.48	.20	-	.23	.31	.51*	.58	.59
	Cross-O	ABM	.78	-	.59	.59	-	.60	.27	-	.26	.34	.61*	.52	.59
Factor	**Study**	**Sample**	**Verbal**			**Quantitative**			**Spatial**			**Perceptual**		**Knowledge**	
			VA	RC	WK	AR	DI	MK	RB	EM	HF	TR	SR/BC	IC	AI
General Factor	Primary	Pilot 0	.49	.42	.50	.72	.74	.65	.37	.30	.31	.34	.61	.27	.10
	Cross-V	Pilot 1	.52	.53	.44	.54	.69	-	.40	.30	.29	.35	.50	.36	.22
	Cross-V	Pilot 2	.46	.50	.39	.74	.72	-	.39	.35	.33	.41	.68	.36	.12
	Cross-V	Pilot 3	.67	.62	.62	.70	.72	.54	-	-	-	-	-	.44	.27
	Cross-O	Pilot 4	.49	-	.38	.61	-	.62	.51	-	.41	.40	.55*	.38	.05
		Pilot *M*	**.53**	**.52**	**.47**	**.66**	**.72**	**.60**	**.42**	**.32**	**.34**	**.38**	**.59**	**.36**	**.15**
	Cross-O	Nav	.50	-	.37	.55	-	.47	.54	-	.43	.38	.59*	.46	.17
	Cross-O	ABM	.48	-	.39	.58	-	.43	.63	-	.60	.43	.60*	.56	.29

Note. Primary = Primary validation study; Cross-V = Cross-validation study; Cross-O = Cross-occupation study