MULTIPLE TEST BATTERIES AS PREDICTORS FOR PILOT PERFORMANCE: A META-ANALYTIC INVESTIGATION

by

Khalid Saif ALMamari

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science in Education

Department of Educational Studies

West Lafayette, Indiana

December 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Anne Traynor, Chair

Department of Educational Studies

Dr. Yukiko Maeda

Department of Educational Studies

Dr. James Greenan

Department of Curriculum and Instruction

Approved by:

Dr. Richard Olenchak

Head of the Graduate Program

ACKNOWLEDGMENTS

First and foremost, I am grateful to The Almighty God for guiding me to complete this research thesis work. “And whatever of blessings and good things you have, it is from Allah” (Quran 16:53). Second, I express my most sincere 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 graduate studies.

I would also like to acknowledge and express my deepest gratitude to my supervisor, Dr. Anne Traynor, for her patience, guidance, support, and encouragement. This thesis would not have been possible without her constructive feedback, thoughtful comments, insights, and knowledge. For their input, guidance, and productive feedback, I also extend my appreciation to my committee member, Dr. Yukiko Maeda, and Dr. James Greenan. For his help, support, and motivation I thank my friend Hamdan Alamri.

To my family, brothers and sisters, I cannot adequately express how much I appreciate your continual encouragement and support. To my sons, Jihad, Albara, Yahya, Azzam, Mohamed, and Saif, I’m truly blessed for being surrounded by such amazing boys in my life.

Last but certainly not least, with my deepest heartfelt appreciation and gratitude, which cannot be fully explained in words, I dedicate this thesis to my wife. Her unstinting patience, support, and encouragement throughout this process were beyond compare. Thank you for making the sacrifices that set me on my academic career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: ALMamari, Khalid, S. MSEd
Institution: Purdue University
Degree Received: December 2018
Title: Multiple Test Batteries as Predictors for Pilot Performance: A Meta-Analytic Investigation
Major Professor: Anne Traynor

A Test Battery (TB) is a measurement method that is designed to assess a variety of ability constructs. The extent to which TB predicts future pilot performance has important implications for both flying organizations and applicants. The primary emphasis in the existing literature has been on scores of individual ability tests, in contrast to the scores of multiple ability tests that are typically indexed by composites derived from TBs. The selection literature lacks a focus on composite scores, and seldom links to the broad cognitive abilities that predominate TBs. The objective of this study was to investigate how the different broad ability constructs of TBs influence their predictive validities for pilot performance. Six ability groups were identified as the most common ability saturations of pilot selection TBs. On the basis of 89 studies and 118 independent samples, a series of meta-analyses were conducted to determine the criterion-related validity of the six categories of TBs for several criterions of pilot performance.

The investigation revealed there was an overall small and positive relationship between TBs and flight performance. The six categories of cognitive ability TBs appeared to be valid predictors of pilot performance, and at least five of them generalize validity across studies and settings. More specifically, three sets of predictor groups were identified according to the magnitude of validity estimates. The highest validity group included *Work Sample* TBs (r=.34), the second highest validity group included TBs of *Acquired Knowledge*, *General Ability*, and

Motor Abilities (*r*=.19, .18, and .17, respectively), and the lowest validity group included TBs of *Perceptual Processing* and *Controlled Attention* (*r*=.14 and .10, respectively).

The results also indicated that there was substantial variability in the effect of cognitive abilities on flight performance, with evidence of moderators operating in most cases. Five potential moderator variables were examined that may moderate the TBs-performance relationship in flying. The analysis for the moderator variable of *the number of tests in the battery* (small battery/large battery), *regularity of TB use in pilot selection* (commonly used/uncommonly used), and *criterion level of measurement* (continuous/ordinal/dichotomous/ contingency table) revealed significant moderating effects on the correlations between flight performance and several types of test batteries. Other moderators related to *year of publication* (1987-1999/2000-2009/2010-2017) and *flying organization* (USAF/US Navy/Another military/Civilian) did not significantly influence the correlations between TBs and flight performance. The implications of the findings for practice are discussed, and recommendations for future research directions are provided.

CHAPTER 1: INTRODUCTION

Test Batteries (TBs) have a long history of use in pilot selection (e.g., Bates, Colwell, King, Siem, & Zelenski, 1997; Carretta, 2000; Damos, 2011; Damos, 1996; Howse, & Damos, 2011; Olson, Walker, & Phillips 2009; Paullin, Katz, Bruskiewicz, Houston, & Damos, 2006; Retzlaff, King, & Callister, 1995; Russell, Reynolds, & Campbell, 1994). A TB is a measurement method that consists of multiple tests/tasks of different abilities (e.g., verbal, quantitative, spatial, psychomotor). Several composite scores are commonly constructed by the sum (or weighted sum) of multiple tests within the battery. Each composite score has different assessment use and utilization (Wong et al., 2012) and varies according to the specific test contents and the higher-order ability factors that explain its structure (Wee, Newman, & Joseph, 2014; e.g., Acquired Knowledge, Perceptual Processing, Motor Abilities, Controlled Attention).

Even with the extensive use of the TBs in the pilot selection, a review of the test validation literature indicates that researchers tend to focus primarily on specific ability test scores and pay less attention to the composite scores derived from multiple TBs. Given the widespread use of TBs in the pilot selection, an understanding of the summary and composite scores produced by a certain battery is critical for interpreting its psychometric properties and for realizing the importance of the underlying constructs (e.g., Aamodt, & Kimbrough, 1985; Bobko, Roth, & Buster, 2007). Therefore, in order to clearly understand the validity of composite scores in the selection context, it is important to investigate the structure of test battery and to highlight the ability domain that influences and explains the derived scores. Meta-analysis researchers who have dealt with other selection methods (e.g., selection interview, work sample tests, situational judgment tests) emphasized that the assessment established from these methods cannot be attributed to one single ability possessed by applicants. Instead, it is a combination of skills and competencies that jointly

contribute to the scores given to applicants, and influence organizations' employment decisions. They have stressed that job-centered methods often do not "cleanly" assess one specific construct, and have suggested, in turn, the concept of construct saturation in order to handle such impurity in measurement (e.g., Christian, Edwards, & Bradley, 2010; Huffcutt, Roth, & McDaniel, 1996; Lievens, & Sackett, 2017; Roth, Bobko , McFarland , Buster, 2008; Roth, Van Iddekinge, Huffcutt, Edison, & Schmit, 2005). Several meta-analyses have been developed to assess the construct saturations of selection methods such as employment interviews (Huffcutt et al., 1996; Roth et al., 2005), work sample tests (Roth et al., 2008), situational judgment tests (Christian et al., 2010), but to date no research effort has been attempted to address the saturation of cognitive abilities dominating selection test batteries.

Indeed, there has been almost no direct meta-analysis that has investigated the validity of the multiple abilities composite scores for pilot performance. This is a critical oversight as understanding the functionality of composite scores in their mixed multidimensional structure provides further evidence of the TBs' construct validity and helps to explain how and why they are related to pilot performance. In the context of pilot selection, in which reliance on the TBs' composite scores has significant consequences, addressing the broad ability constructs that saturate TBs contributes to our understanding of these constructs and of the relative role they play in predicting pilot performance. Hence, this study was designed to meta-analyze the criterion-related validity of TBs (i.e., composite scores) for pilot performance, after developing a categorization schema for TBs centered on the broad constructs of abilities.

1.1 Statement of the Problem

The examination of summary and composite scores resulting from batteries of two or more tests is essential and has both theoretical and practical implications. It is these composited scores, rather than the individual test scores, that are most important for recruiting offices and that are used for making the overall assessments and the selection decisions (e.g., Cowan, Barrett, & Wegner, 1990). The TBs' composite scores have more potential for representing the individuals' latent abilities, and therefore warrant the same, if not more, attention as that given to the individual ability scores. Gibbons et al., (2012) demonstrated at least four advantages of composite measures over individual measures, which include the following (1) they may be more powerful, (2) they increase measurement precision, (3) they help to avoid the individual features of a particular test that may capitalize on chance, and (4) they limit the number of statistical tests required compared to analyzing each component separately.

From a meta-analysis perspective, the lack of emphasis placed on the criterion-related validity of composite scores may be understood by considering the typical complex structure of the aggregated scores. The ability tests incorporated within the TBs vary substantially, such that the attempt to compile them into one broad class of predictor would result in a mixture of oranges and apples that good meta-analysis makes every effort to avoid (Rosenthal & DiMatteo, 2001). Similarly, an attempt to group them into a few broad categories to facilitate data accumulation is not always viable and can be extremely challenging. To resolve this categorization dilemma, which could impede the attempt to investigate the validity of composite scores meta-analytically, recent developments in intelligence models and recent investigations in selection methods may provide a frame of reference to support the desired investigation. The abilities taxonomy proposed by the Cattell–Horn–Carroll (CHC) theory of cognitive abilities incorporates three strata of abilities that

cover the ability domains extensively and expansively (McGrew, 2005; Schneider & McGrew, 2012). This theory is considered the most comprehensive and empirically supported psychometric model of the structure of cognitive and academic abilities to date (Alfonso, Flanagan, & Radwan, 2005).

In the CHC theory, cognitive abilities are described as the fundamental construction of human intelligence and performance processes. The CHC three-stratum model is an incorporation of the two most prominent models in human cognitive abilities: The Cattell–Horn model ($G_{fluid\ reasoning}$ -$G_{crystallized\ intelligence}$ theory (G_f-G_c); Cattell, 1943; Horn, 1968; Horn & Cattell, 1966) and Carroll's three-stratum model (1993). The primary distinction between the two models is the inclusion of a general ability factor, *g*, at Stratum III for the Carroll model but not the Cattell–Horn model. In CHC theory, general intelligence (*g*) is the third stratum of the model (stratum III), followed by 16 broad cognitive abilities at the second stratum (stratum II), and 80 or more specific cognitive abilities at the first stratum (stratum I; McGrew, 2009). A conceptual model of CHC theory organizing the broad cognitive abilities into even broader abilities was found a good fit for the categorization of ability batteries required for this study. Likewise, from the literature of meta-analysis in job selection methods it was possible to draw on the idea of construct saturation, which refers to the extent to which a specific construct influences (or saturates) complex measures. The term "saturation" was used by Lubinski and Dawes (1992, p. 28) to denote how a given construct (e.g., *g*-saturation, motor abilities-saturation, perceptual processing-saturation) have an effect on a complex, multidimensional measure. Because construct saturation can affect validity and sub-group differences, it can be viewed as a mediator of the relationship between selection method factors and these outcomes (Lievens & Sackett, 2017). The importance of such an adaptable approach is that it can accommodate the composite scores of heterogeneous abilities tests, and can

tolerate the mixed and diverse contents of these scores. This concept supported the investigation of other selection methods, such as saturation of employment interviews with cognitive ability (Huffcutt et al., 1996), saturation of structured interviews with personality (Roth et al., 2005), saturation of work samples (Roth et al., 2008), and situational judgment tests (Christian et al., 2010) with different ability constructs.

With this study, I seek to close this research gap by integrating the body of knowledge available for the criterion-related validity of TBs versus the different outcomes of pilot performance. Criterion-related validity refers to the evidence of validity collected for a particular test by assessing its score's correlation with a score on an external criterion. Two types of criterion-related validity are commonly assessed: predictive and concurrent. Predictive validity is defined as "the degree to which test scores predict criterion measurements that will be made at some point in the future" (Crocker & Algina, 1986, p.224). Concurrent validity is established by comparing (i.e., correlating) the test scores on an instrument with scores on another (criterion) measure that is measured concurrently in the same subjects (Kimberlin & Winterstein, 2008). I also extend with this study the earlier efforts initiated by two comprehensive meta-analyses (Hunter & Burke, 1992; 1994; Martinussen, 1996) that focused primarily on ability-specific tests used in pilot selection (e.g., verbal, quantitative, spatial, information processing). Both studies accumulated the TBs' composite scores under one category group, without much consideration on the orientation of ability tests contributed to the composites, or an appraisal of their contents' saturation with predominant ability domains. Hence, the research problem pursued in this work is to assess meta-analytically the relationships between scores of TBs composites and scores of flights performance rating to understand the contribution of cognitive abilities in flight performance.

1.2 Purpose of the Study

This study responds to the calls for Industrial and Organizational psychology researchers to reconnect with the science of mental abilities and measurement theory to gain a better understanding of how constructs within the intelligence nexus manifest in the context of work (Reeve, Scherbaum, & Goldstein, 2015). The primary objective was to conduct a meta-analytic investigation of the validity of test batteries for predicting pilot performance. To achieve this objective, a functional approach was used for classifying the TBs, for both practical and theoretical reasons. Informed by the CHC theory's conceptual model of broad abilities (Schneider & McGrew, 2012), four categories were constructed to correspond to the TBs that were saturated with the broad abilities of *Acquired Knowledge*, *Perceptual Processing*, *Motor Abilities,* and *Controlled Attention*. Two more categories were added, one of which corresponded to the TBs saturated with *General Ability* and the other of which correspond to the *Work Sample* mode of TBs, which is administered using a flight simulator. The broad ability-based categories were found to cover a large portion of the available test batteries commonly used in pilot selection and assessment. From a practical perspective, this framework provided a useful categorization that can be easily understood and applied, although it may not be exhaustive. The classification of TBs into six broad ability constructs aided the accumulation process of the widely-differing TBs and enabled their subsequent analyses. From a theoretical perspective, connecting the TBs with CHC theory-based model helps to emphasize more explicit the broad constructs of cognitive abilities, which inform our perception on the validities of these abilities as predictors for pilot performance. If the measurements provided by TBs upon selection are found to predict the subsequent performance, this would be relevant information for both potential students and recruiting offices of flying organizations. The assessment resulted from the TBs could be used in decisions about application

and admission, and it could also be used in considering modifications that may be required to improve student success. The findings could potentially assist in the future development of TBs used for screening pilots' applicants. Such a result could reduce attrition of students from the flight training program, and increase competency and safety of trainees.

Accordingly, the present study meta-analyzed the criterion-related validity of the six identified broad ability construct saturations of TBs, namely, Acquired Knowledge, Perceptual Processing, Motor Abilities, Controlled Attention, General Ability, and Work Sample for four specific criteria of pilot performance (i.e., flying performance rating, graduate and attrite training, academic performance grade, flight simulator performance rating) and one overall criterion (the best index of flying performance presented by each study). In order to extend the previous efforts without repetition, the search for studies was narrowed down to a few years before the publication of the pilot selection test meta-analyses of Hunter and Burke (1994) and Martinussen (1996), more specifically, from 1987 to the present.

1.3 Research Hypothesis and Questions

This study investigated the criterion-related validity of six categories of TB composite scores within the context of pilot selection and assessment using the meta-analysis technique. Previous validation research was used to guide the development of the research hypothesis and questions (more information is provided in the next chapter). The following hypotheses were posited to direct the investigation:

Hypothesis 1. The six ability saturations of the test batteries (Acquired knowledge, Perceptual Processing, Motor Abilities, Controlled Attention, General Ability, Work Sample) will show small to moderate mean correlations for predicting pilot performance across four specific criteria (flying

performance rating, graduate/attrite training, academic performance grade, flight simulator performance rating), and they will generalize validity across samples and settings.

Hypothesis 2. The six ability saturations of the test batteries (Acquired knowledge, Perceptual Processing, Motor Abilities, Controlled Attention, General Ability, Work Sample) will show small to moderate mean correlations for predicting the overall criterion of pilot performance (the best index of flying performance presented by each study), and they will generalize validity across samples and settings.

In order to compare the six broad ability saturations of TBs and the four criterions of pilot performance, the following research questions were investigated:

Question 1. Among the six ability saturations of the test batteries, which is the best predictor for each specific performance criterion?

Question 2. Of the six ability saturations of the test batteries, which is the best predictor for the overall criterion of pilot performance?

In addition, five variables were identified as potential moderators for the associations between predictors and criteria: number of tests in the battery, regularity of TB use in pilot selection, year of publication, flying organization, and criterion level of measurement. In the context of meta-analysis, a moderator variable can be any situational feature or human attribute that differentiates between subgroups within the sample (Levine, Spector, Menon, & Narayanan, 1996). Moderators analyses help to explain differences in the strength or direction of observed relationships between the primary variables of interest (Steel & Kammeyer-Mueller, 2002). The search for moderators is warranted when the variability around the mean validity is found to be significant, indicating substantial heterogeneity (Tett, Hundley, & Christiansen, 2017). Therefore, the following hypotheses and questions about potential moderating variables were formulated:

Hypothesis 3. Large test batteries with five or more tests will predict pilot performance better than small test batteries with fewer than five tests.

Hypothesis 4. Test batteries commonly used in pilot selection will predict pilot performance better than test batteries that are used less often in pilot selection.

Hypothesis 5. Validity estimates of the six test batteries will tend to decrease in more recent publications (1987-1999, 2000-2009, 2010-2017).

For the remaining two moderators, no specific hypotheses were posited due to inadequate evidence. Instead, the following questions are posed:

Question 3. Do the validities of test batteries vary as a function of the flying organization (USAF, US Navy, Another Military, Civilian)?

Question 4. Do validities of test batteries vary as a function of performance criterion level of measurement (Continuous, Ordinal, Dichotomous, Contingency table)?

1.4 Operational Definitions of Test Battery Categories

This study focused on the relationship between the six broad ability saturations of TBs and several criteria of pilot performance. A clear definition of predictor terminology is critical to explain their usage in the study. Thus, the following describes the operational criteria exploited for categorizing the collected TBs. A more theoretical basis for the classification schema can be found in the succeeding chapter.

1.4.1 Test Battery Saturated with Acquired Knowledge. It is characterized by five main features (a) Predominated by verbal tests, (b) Primarily uses verbal and quantitative cognitive tests, (c) Often includes aviation or general knowledge tests, (d) May contain fewer spatial and perceptual-cognitive tests, (e) Commonly has a paper-and-pencil format.

1.4.2 Test Battery Saturated with Perceptual Processing. It is characterized by five main features (a) Predominated by nonverbal tests, (b) Primarily uses spatial and perceptual-cognitive tests, (c) Often includes visualization and memory tasks, (d) May include fewer verbal and quantitative cognitive tests, (e) Commonly administered via computer.

1.4.3 Test Battery Saturated with Motor Abilities. It is characterized by 5 main features (a) Primarily focused on psychomotor coordination tests, (b) Dominated by tasks requiring movement of fingers, hands, and legs, (c) Often includes tasks of compensatory tracking and reaction time, (d) May involve cognitive tests of different abilities, (e) Commonly performed using advanced computer or other specialized apparatus.

1.4.4 Test Battery Saturated with Controlled Attention. It is characterized by 5 main features (a) *Always* involves dual tasks (or multitasking), (b) Dominated by dual tasks requiring controlled attention and sensory processing, (c) Often includes working memory, time-sharing, and shifting attention tasks, (d) May include psychomotor dual tasks (tracking with joystick and foot pedals), dual cognitive tasks (memory, math), or mixed motor-perception dual tasks (piloting and listening), (e) Commonly conducted using advanced computer or specialized apparatus.

1.4.5 Test Battery Saturated with General Abilities. It is characterized by 5 main features (a) Includes mixed types of tests (verbal, nonverbal), (b) No clear domination of specific test orientations, (c) Often integrates and incorporates a large number of tests representing multiple abilities, (d) May provide intelligence quotient or known for its *g*-saturation (general intelligence), (e) Can be administered in a variety of formats (paper and pencil, computer-based, apparatus-based).

1.4.6 Test Battery Saturated with Work Sample. It is characterized by 5 main features (a) *Always* involve a flight simulator, (b) Dominated by tasks simulating actual flying experience, (c)

Often requires completing a number of simulated flight maneuvers, (d) May utilize a whole-body motion simulator simulating flying a small single-engine aircraft, (e) Commonly performed using an advanced motion flight simulator.

CHAPTER 2. LITERATURE REVIEW

This chapter covers topics related to aviation psychology such as aspects of flight training programs, suggested models for pilot selection, pilot-related psychometric meta-analysis, individual ability tests versus multiple ability tests, categorization of TBs' composite scores, and considerations for the proposed classification schema.

2.1 Flight Training Program

Psychological selection tests as predictors of pilot performance have been a subject of considerable attention and extensive research effort (e.g., Bates et al., 1997; Howse & Damos, 2011; Johnston, 1996; Russell et al., 1994; Weissmuller & Damos, 2014). The expanding interest in pilot assessment may be linked to historical, economic, theoretical, and practical reasons. The training of pilots is long, challenging, and extremely expensive (Carretta et al., 2014; King et al., 2012) at a cost exceeding $900,000 per aviator (Ostoin, 2007) and ranging from $500 to over $3,000 per flight hour (Hunter & Burke, 1992). The U.S. Navy spends up to $880,000 simply to provide a student aviator the training required to start flying the type of aircraft he or she is assigned to, followed by additional costs for training specific to that platform (Olson et al., 2009). In addition to economic factors, Griffin and Koonce (1996) expected that the cost of training pilots would continue to escalate as a result of the technological enhancement of aircraft. Despite the rigorous selection requirements, training attrition of pilot students continues to be a significant concern, with a rate exceeding 25% (Hunter & Burke, 1994) and the average cost for each failure ranging from $50,000 to $80,000 (Hunter, 1989; Siem, Carretta, & Mercatante, 1988), reaching as high as $1 million in some services (Helm & Reid, 2003). This can be expected due to the strict timetable of flight training and the demanding training environment. Attrition is a waste of an

opportunity that another applicant could have used and a loss of an investment in training slot costs from which no return is expected (approximately $750,000 for a USAF trainee; Williams, 2009). Most importantly, if the attrition rate exceeds a certain level, a shortage of pilots will occur, which will affect the readiness of flying organizations (Lynch, 1991).

2.2 Suggested Models for Pilot Selection

A well-established model for pilot selection is highly sought-after as it would help to solve the problem of high attrition rates commonly experienced in training programs, as well as, in the long term, contribute to a more effective and resilient organization (Martinussen & Hunter, 2009, p. 73). The ongoing validation studies of pilot selection tests show that some abilities are better predictors of pilot performance than others. Carretta and Ree (1996) asserted that general intelligence is by far the best predictor of pilot training success. Additionally, a model for selection containing intelligence tests, psychomotor tests, personality tests, and information processing tests has been suggested, with validity coefficients ranging from .20 to .40 (Carretta, 1992). Recently, the examination of neurocognitive test batteries for USAF pilot trainees revealed that general cognitive ability was the main predictor of pilot performance, and there was little evidence that any specific cognitive variable was more important than any other (King et al., 2012). Other evidence suggests that perceptual speed, quantitative, and aviation knowledge had the highest validity coefficients across different criteria (Johnson, Barron, Rose, & Carretta, 2017). In the comprehensive narrative review of pilot selection, Paullin et al. (2006) recommended that the U.S. Army should focus on measures of cognitive abilities such as spatial ability, mechanical reasoning, verbal ability, numerical reasoning, and perceptual speed, as well as a measure of motivation to become an aviator. On the whole, although there is some overlap in the types of cognitive abilities

that are proposed to be relevant to pilot selection, there is not a single widely agreed-upon model explaining their relationships.

2.3 Pilot-related Meta-Analysis

Given this lack of consensus in the primary validation studies, meta-analyses of selection tests have attracted growing interest as a way to improve the prediction of success in training programs and job performance. Such analyses integrate findings and statistics of prior research to estimate the population mean correlation (Glass, 1977) and also helps to improve the statistical power associated with the predictor-criterion relationship (e.g., Cohn & Becker, 2003). The validity generalization approach to meta-analysis (Schmidt & Hunter, 2015) has been the most widely used approach in organizational psychology (Kepes, McDaniel, Brannick, & Banks, 2013). It serves to establish whether a particular psychological construct, test, or measure has validity in predicting job performance, irrespective of situation or setting (DeGeest & Schmidt, 2010). Lately, the term has been referred to as "psychometric meta-analysis" in recognition of the fact that the methods correct for the biasing effects of statistical artifacts such as unreliability and range restriction, whereas other meta-analytic methods typically correct for only sampling error (Ones, Viswesvaran, & Schmidt, 2017). These research synthesis methods have had a broad impact in the field of industrial/organizational psychology and the related disciplines of human resources management (HRM) and organizational behavior (DeGeest & Schmidt, 2010).

Oddly, even with the increasing interest in criterion-related meta-analyses, pilot-related meta-analyses are still limited and cover only a few aspects of human factors in the flying profession. Out of few meta-analyses, only two reviews have studied the selection tests comprehensively and expansively (Hunter & Burke, 1994; Martinussen, 1996). The rest of the

investigations were limited in scope and focused on specific domains such as personality (Campbell, Castaneda, & Pulos, 2009; Castaneda, 2007) and multitasking (Damos, 1993), or focused on specific selection test battery such as those used by the Norwegian Air Force (Martinussen & Torjussen, 1998), the U.S Air Force (Lynch, 1991), or the U.K Royal Army (Burke, Hobson, & Linsky, 1997). Other meta-analyses were designed to evaluate the effect sizes of flight simulator training (Hays, Jacobs, Prince, & Salas, 1992), simulator platform motion (Vaden & Hall, 2005), whole-body flight simulator motion (De Winter, Dodou, & Mulder, 2012), and crew resource management (O'Connor et al., 2008).

The results from pilot-related meta-analyses revealed important findings supporting or contradicting several long-held assumptions. For example, in contrast to the emphases given to general intelligence as the best stand-alone predictor of pilot performance (Carretta & Ree, 1996), Hunter and Burke (1994) and Martinussen (1996) came to the overall conclusion that tests measuring general intelligence have lower predictive validity than tests measuring certain specific cognitive and psychomotor abilities and biographical information. As regards personality constructs, Campbell et al. (2009) showed that the neuroticism and extroversion dimensions of the five-factor model of personality, as well as the specific facet of anxiety, were valid predictors for pilot performance. On the topic of multitasking, Damos (1993) found that multi-tasking ability is more predictive than the single-tasking ability for pilot performance. In their narrative review, Paullin et al. (2006) highlighted the role of abilities other than intelligence, such as psychomotor ability and information processing skills, as constructs believed to add incremental validity beyond that achieved by general intelligence.

Relatedly, Hays et al. (1992) examined the effectiveness of flight simulator training, showing that the use of simulators produced improvements in training for jet pilots. Another meta-

analysis conducted to evaluate the effect of simulator platform motion on pilot training transfer suggested that simulator motion has a small, positive impact on pilot training transfer (Vaden & Hall, 2005). A meta-analysis of training effectiveness of whole-body flight simulators revealed that motion appears essential for flight-naive individuals learning tasks, but not for experts learning fixed-wing aircraft maneuvering tasks (De Winter et al., 2012). Crew resource management (CRM) was also subjected to meta-analysis to evaluate its effectiveness in training. The findings generally supported the effectiveness of CRM training, with substantial effects observed on the participants' attitudes and behaviors and medium effects on their knowledge (O'Connor et al., 2008). Despite the meta-analyses above, the majority were narrow in scope and focused on a specific area of aviation components. Comprehensive meta-analyses were only made available by Hunter and Burke (1994) and Martinussen (1996), both of which were conducted nearly 25 years ago. An updated meta-analysis of the new literature is vital, especially for score types that have not been adequately covered in previous reviews such as the composite scores of test batteries.

2.4 Comprehensive Psychometric Meta-Analysis

During the past 25 years, there have been two comprehensive meta-analyses published relating a wide range of cognitive and non-cognitive test scores to flight performance criteria (Hunter & Burke, 1994; Martinussen, 1996). In addition to these two studies, there exists a small scale meta-analysis focused exclusively on pilot selection tests of Norwegian Air Force (Martinussen & Torjussen, 1998). The summary findings of the three meta-analyses are illustrated in Figures 1 – 3.

Hunter and Burke's (1994) meta-analysis reviewed 68 validation studies conducted between 1940 and 1990. The combined observed correlations between sixteen categories of

predictor test scores and various flight training outcome criteria were investigated for their predictive validity. Overall, the 16 types of test scores appeared valid predictors of flight training success. Mean correlations of the predictors with outcomes ranged from .20 to .34 for eight categories, from .10 to .19 for six categories, and .06 for one category. Figure 1 shows the details of these predictor groups (age group was not included; *r*=-.1). Ten of the 16 categories had limited capacity to be generalized across samples and settings; the other 6 had no capacity for generalization (general intelligence, verbal skills, fine motor ability, age, education, and personality).

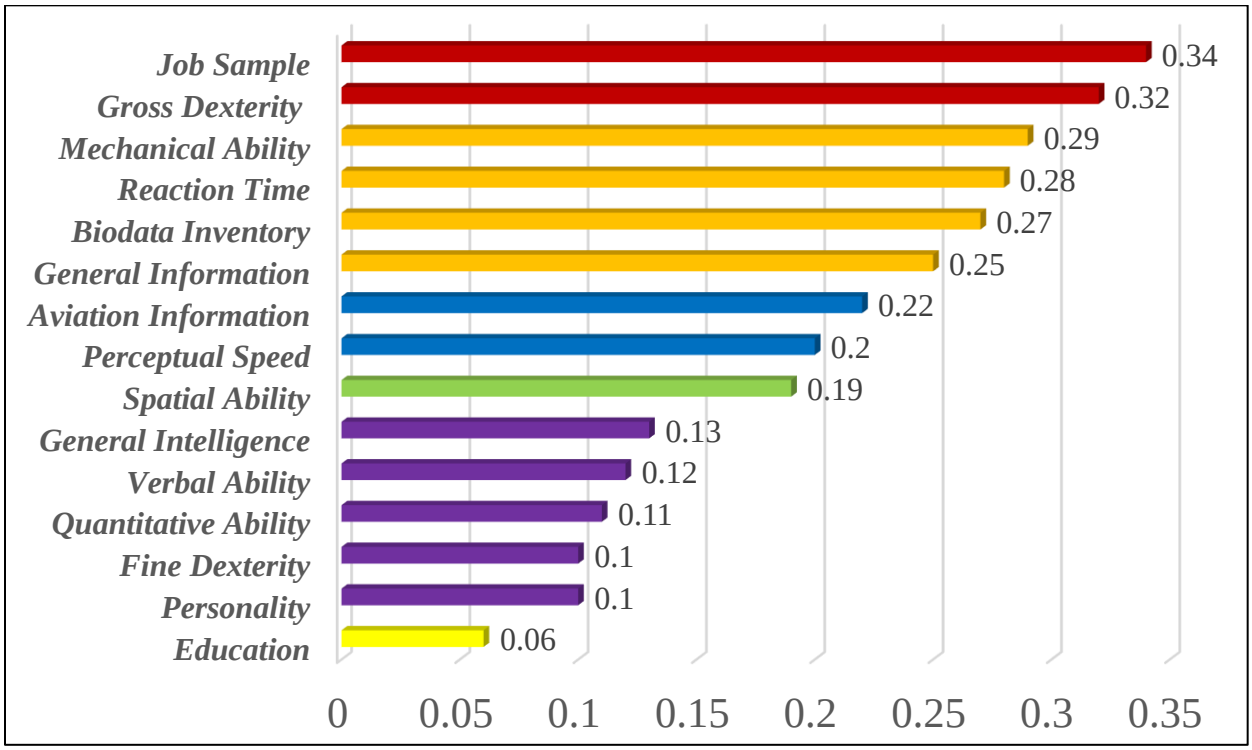

Figure 1. Hunter & Burke's (1994) meta-analysis

Martinussen (1996) published a newer meta-analysis two year after Hunter and Burke's (1994) study, by collecting studies from 11 different nations. Nine group of predictors were constructed out of 66 independent samples from 50 studies. Figure 2 includes the magnitude of effect sizes. Results indicated that flight training experience (r=.3) was the best predictor for predicting flight training success, followed by three cognitive ability groups with mean validities of .24 for each (combined index of cognitive abilities, aviation information, and psychomotor/information processing), whereas intelligence tests, academic, and personality tests had the lowest mean validities.

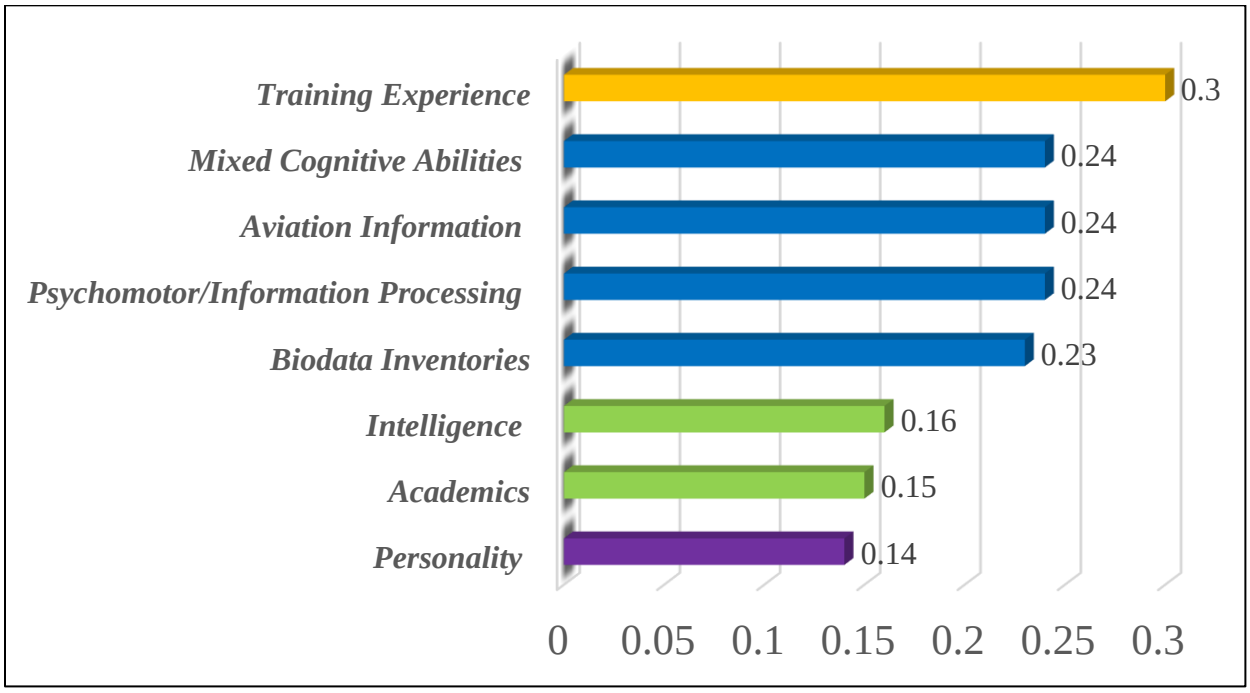

Figure 2. Martinussen's (1996) meta-analysis

Martinussen and Torjussen's (1998) small-scale meta-analysis evaluated the utility of the Norwegian Air Force selection test battery as a predictor of flight training performance. Based on five studies, the 20 tests built-in within the selection battery, were meta-analyzed in relation to criteria of pilot performance. It was determined that the best three predictors of success in flight training performance were instrument comprehension (*r*=.26), aviation information (*r*=.21), and mechanical principles (*r*=.19). A notable finding was that three tests yielded negative mean validities (Rotating Patterns, Paper Fonning, Numbers). Figure 3 charts results of the 20 meta-analyses conducted in this study.

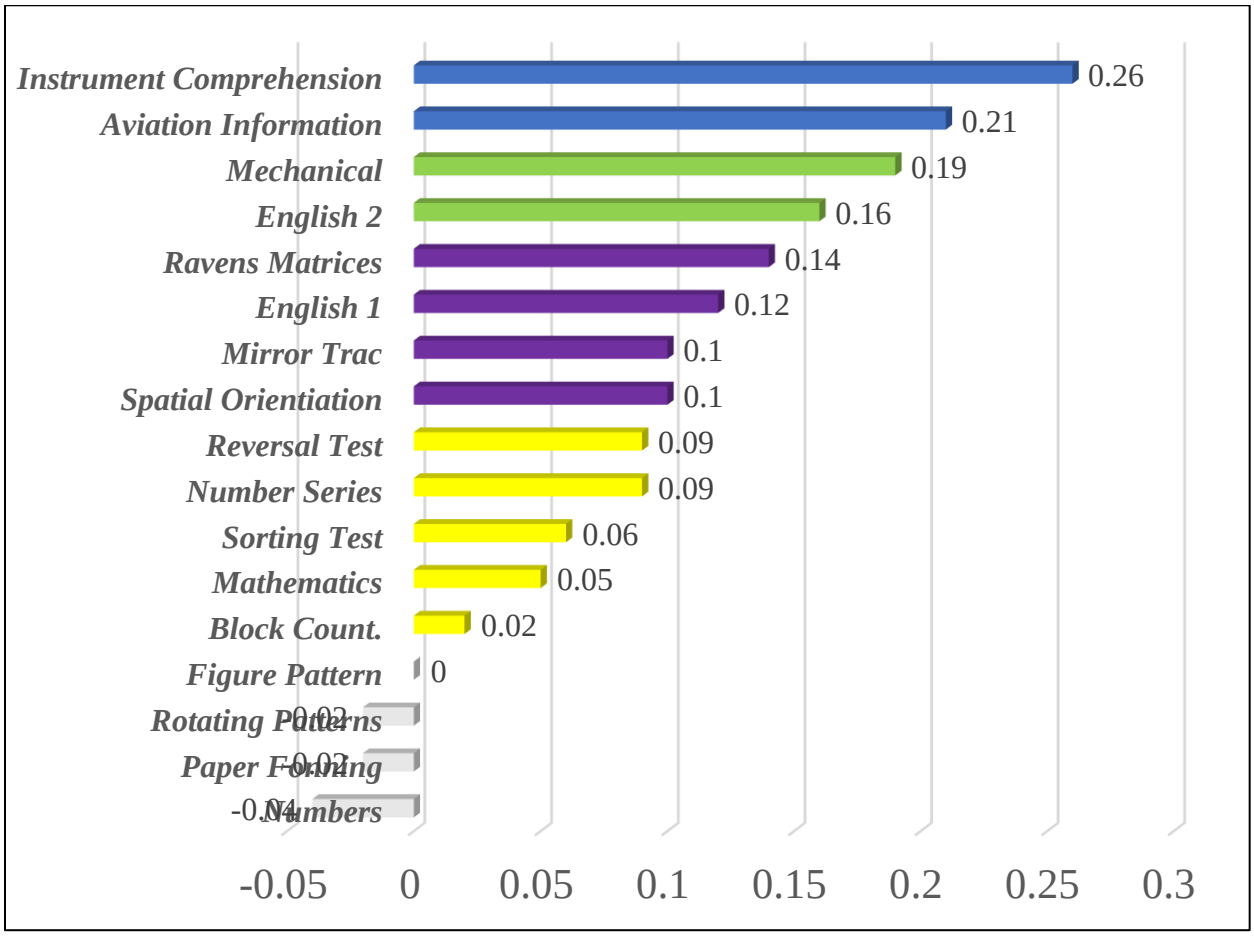

Figure 3. Martinussen and Torjussen's (1998) small-scale meta-analysis

2.5 Individual Ability Tests versus Multiple Ability Tests

The comprehensive meta-analyses of Hunter and Burke (1994) and Martinussen (1996) have contributed to our understanding of the individual differences in piloting ability and aided our perception of the most central abilities for pilots. This has informed selection policy and planning (e.g., Lochner & Nienhaus, 2016; Paullin et al., 2006) and motivated test publishers to design batteries based on the results of these studies (Kokorian et al., 2004). Both research articles are among the most cited articles in aviation psychology and have been exhaustively described and demonstrated (e.g., Damos, 2011; Martinussen & Hunter, 2009, p. 88; McFarland, 2017; Paullin et al., 2006; Reinhart, 1998). This is not surprising since there was (and is) no other comprehensive meta-analytic examination of the predictive validity of psychological tests for pilot performance.

Nevertheless, the primary focus of both investigations was given to the individual constructs of abilities (e.g., verbal, quantitative, spatial) as they are measured with individual tests or group of tests designed for that specific construct. In contrast, the overall indexes of mixed abilities (i.e., composite scores) derived from multiple tests were not adequately addressed, although they often represent the final product of assessment batteries (Bobko et al., 2007). In practice, psychological tests are typically administered as part of a battery, not in isolation. The use of multiple tests rather than single tests for overall assessment purposes has frequently been advocated. For example, in the case of Wechsler's intelligence scale, Zachary (1990) stressed that the subtests do not exist to measure any specific ability in isolation but to assess something that will emerge from the individual's performance as a whole.

Similarly, in a critique of Wechsler theory and practice, McDermott, Fantuzzo, and Glutting (1990) emphasized the role of the multiple test scores and insisted, "Just say no to subtest analysis." It is

apparent that the cognitive abilities manifestations are entwined such that the attempt to separate their assessment into individual test scores may understate the measurement of the overall cognitive function. However, integrating the assessments from multiple ability tests into a single overall composite score may match the ability interconnections more appropriately. Given the complex abilities' interplay in piloting, Bates et al. (1997) stated that "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." Nonetheless, when considering the structure of the cognitive system, only a limited number of tests are ability-specific, although they are also dependent on the integration of other parts of the cognitive system (e.g., Crane et al., 2008; Embretson, 1998; Heaton et al., 2014; Hornby, 2003). Hence, the investigation of summary and composite scores stemming from the multiple tests batteries is crucial and has important implications.

2.6 One Single Category for Composite Scores

From the limited analysis allocated to composite scores of mixed abilities in Hunter and Burke (1994) and Martinussen (1996), it was clear that the mean validity estimates of these predictor categories were among the best predictor groups, even higher than the intelligence test category, which is arguably the best predictor for performance. However, Hunter and Burke (1994) disregarded reporting the meta-analyzed validity of the combined index in their published paper although they had considered it in the technical report upon which the published article was based. The results from the technical report (Hunter & Burke, 1992) found a validity of .19 (N= 34) for a group of predictors designated as "Composites/ Batteries" for the scores derived from the combination of a number of separate tests. Even so, it appears that they added this index without

much interest, as they stated that “the categories of Composite/Battery and Other are included in this analysis solely for the sake of completeness of reporting.” Martinussen (1996) meta-analysis, on the other hand, designated a special category for “Combined index” for the combination of several tests, usually both cognitive and psychomotor, for which the correlations between the subtests and the criteria were not reported. The highest mean validity (corrected for dichotomization) in the meta-analysis was obtained from this category group (r= .31, N= 14). Apart from a very broad definition provided for this index, no clear information was offered about the content that shaped the combinations of cognitive and psychomotor abilities.

The lack of focus on the predictive validity of the composite scores may be justified given their typical complex structure. A composite score is typically formed by a joint contribution of several ability tests and, in many cases, represents the best combination of tests in the battery for measuring a certain latent construct. Given the multiplicity and diversity of tests involved, the attempt to group TBs’ composite scores into a single class of predictor (as was the case in the two meta-analyses) may be not meaningful and may reveal little about the construct saturating the combined scores. Equally, the attempt to group them into multiple classes of predictors is a challenging exercise and necessitates a thorough understanding of the specific and broad structure of TBs. It also requires a well-defined classification schema that has room for the variety of ability constructs that emerged from TBs’ summary scores. From another perspective, there is also a need for an acceptable procedure that accommodates the heterogeneity structure of the aggregated scores and takes into account the primary construct saturation of the scores along with any secondary constructs that play parts in the background.

The recent development in intelligence theory and some recent practices in the meta-analysis of selection methods may provide solutions for the problems related to TBs categorization

and construct impurity. For that reason, a compatible framework was developed to support the desired investigation. The broader abilities taxonomy proposed by the Cattell–Horn–Carroll (CHC) theory of cognitive abilities (McGrew, 2005; Schneider & McGrew, 2012) were used for categorizing test batteries. In addition, the construct saturation concept endorsed by some meta-analysis researchers in selection methods was found to be a conceivable showcase for the possibility of investigating constructs with complex structures (e.g., Christian et al., 2010; Huffcutt et al., 1996; Roth et al., 2005, 2008).

2.7 Six Categories for Composite Scores

CHC theory identifies *g* at the top of the hierarchy (stratum III), 16 broad abilities below *g* (stratum II), and around 80 narrow abilities (stratum I) nested within the broad abilities. There are intermediate categories between the three strata in several places, and some parts of the taxonomy are more settled than others (Schneider & Flanagan, 2015). Due to the complexity of the taxonomy and the increasing number of abilities added to the model, the need for an overarching framework with which to understand CHC theory as a whole became critical. One proposed model is the higher-order groupings of the broad abilities in CHC theory (Schneider, Mayer, & Newman, 2016; Schneider & McGrew, 2012; Schneider & Newman, 2015). Reducing the complexity, conceptual economy, and clustering abilities of common features were some explanations given for the formation of this model. As seen in Figure 4, the model organized through four "big" higher-order abilities (Acquired knowledge, Perceptual Processing, Motor Abilities, Controlled Attention) integrates the broad abilities defined by the theory (e.g., verbal comprehension, fluid reasoning, visual-spatial, psychomotor abilities). Figure 4 also includes two distinct factors of abilities: "level" factors and "speed" factors. At the level factors, abilities are defined by the difficulty of the task while defined at speed factors by the rate of completing the tasks. Because of the speed-

accuracy trade-offs in most tasks, the distinction between level and speed is not necessarily a true dichotomy (Schneider & Flanagan, 2015).

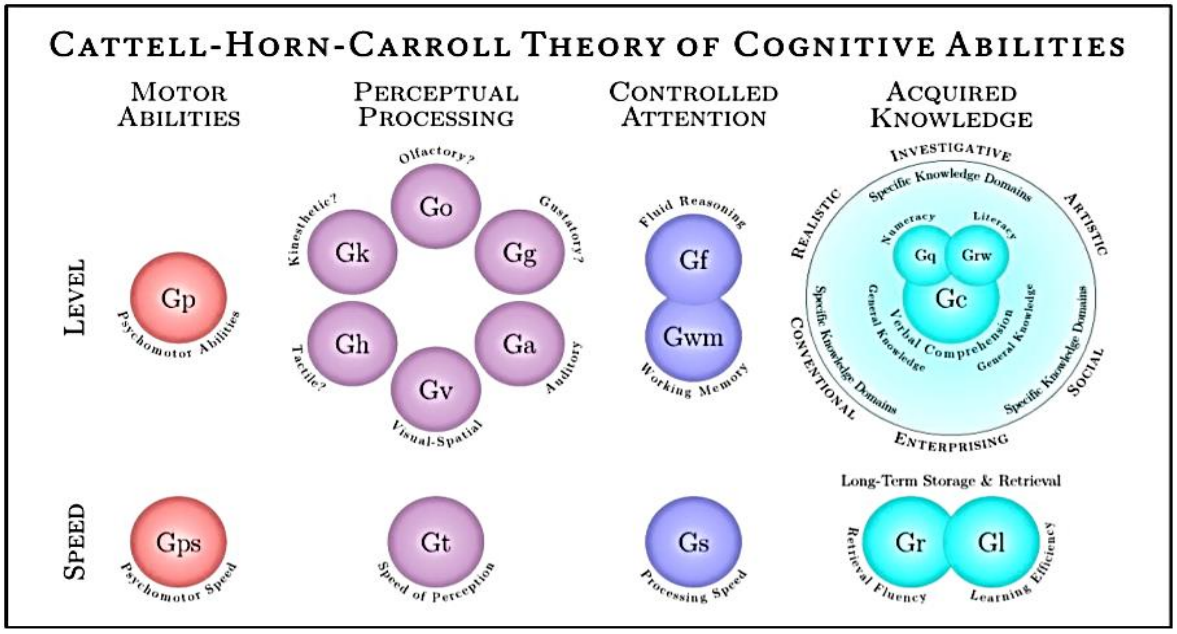

Figure 4. Conceptual grouping of broad abilities in the Cattell-Horn-Carroll (CHC) theory (Reproduced with permission from Schneider, Mayer, & Newman, 2016).

Based on this conceptual model suggested by CHC theorists, the four broader abilities were found to satisfy a large portion of the classification need for the TBs in this study. This taxonomy of cognitive abilities provides a frame of reference, terminology, and a map to position ability variables (i.e., composite scores) found in the literature, thereby facilitating construct categorization and accumulation. Accordingly, four categories of TBs' composite scores corresponding to the four broad abilities were identified. In addition, a fifth category was constructed to correspond to the General Ability, which appears in most intelligence theories, including CHC theory. This group category was necessary to accommodate test batteries with *g*-

saturation, and those integrating a large number of mixed-type tests. Moreover, a sixth category related to work sample was also considered to accommodate the approach of pilot selection employed by some flying organizations that use flight simulators as a form of TB. Given the variety of abilities and traits that contribute to the overall performance in work sample tests or flight simulator (e.g., cognitive, psychomotor, knowledge, personality), this category group was termed by its main character as a Work Sample, irrespective of the broad cognitive or psychological factors that explain and influence performance. Therefore, six broad construct saturations formed the basis for mapping and grouping the multiple tests/tasks batteries collected for the meta-analyses: (a) Acquired Knowledge, (b) Perceptual Processing, (c) Motor Abilities, (d) Controlled Attention, (e) General Ability, and (f) Work Sample. The operational criteria used for classifying the TBs was described in the previous chapter. A brief description of the main features of each category of TBs is presented below.

2.7.1 TBs Saturated with Acquired Knowledge

In CHC theory, the broader ability domain of Acquired Knowledge includes broad abilities such as verbal comprehension, numeracy, literacy, general knowledge, and domain-specific knowledge. It corresponds to Cattell's crystallized intelligence and Horn's expertise abilities domains (Schneider et al., 2016). Knowledge domains cover abilities that are highly valued by one's culture as well as those acquired because of particular interests or vocational requirements (Schneider & McGrew, 2012). Many selection test batteries include tests that can be categorized under this broad factor. In fact, the traditional test batteries for pilots are typical of this type as they emphasize verbal and quantitative abilities as well as specific knowledge such as aviation knowledge and instrument comprehension. Examples of TBs saturated with Acquired Knowledge

include the Air Force Officer Qualifying Test (AFOQT; pilot composite; Drasgow, Nye, Carretta, & Ree, 2010), the Aviation Selection Test Battery (ASTB; PFAR composite; Lopez & Denton, 2011), and the Multidimensional Aptitude Battery-II (MAB-II; verbal IQ composite; Carretta, King, Ree, Teachout, & Barto, 2016). As this group represents the most common form of selection test batteries, it was anticipated that it would be an important predictor of pilot performance.

2.7.2 TBs Saturated with Perceptual Processing

This broader ability includes two essential skills for pilots: visual-spatial and auditory processing abilities. Also, it contains kinesthetic, tactile, olfactory, and gustatory abilities that have yet to be supported as distinct abilities (Schneider & McGrew, 2012). Perceptual Processing, along with Motor Abilities, both correspond to Cattell's broad ability of provincial powers (Schneider et al., 2016). The importance of these abilities may be realized by inspecting the available test batteries that are rarely absent from tests measuring spatial ability or information processing (e.g., Carretta, 1987, 1988; Morrison, 1988). Applicable examples for TBs saturated with Perceptual Processing include the selection tests of the German Aerospace Center (DLR; composite of 7 cognitive abilities; Zierke, 2014) and Norwegian Air Force (multiple composites from 20 psychological tests; Martinussen & Torjussen, 1998). Additionally, the recently validated neurocognitive assessment batteries for USAF pilots offer good examples of this ability saturation, namely the MAB-II (performance IQ; Carretta et al., 2016), the Microcog (King et al., 2013), and the Cogscreen (King, Barto, Ree, Teachout, & Retzlaff, 2011). Several validation studies supported the effectiveness of TBs grouped into this category; thus, the expectation was that it would show positive relations with outcome criteria.

2.7.3 TBs Saturated with Motor Abilities

According to CHC theory, the broad domain of Motor Abilities is composed of psychomotor abilities and psychomotor speed. The ability portion of the motor factor refers to the capacity to rapidly and fluently perform body motor movements (movement of fingers, hands, legs, etc.) independent of cognitive control, while the speed portion refers to the ability to perform body motor movements with precision, coordination, or strength (McGrew & Evans, 2004). In aviation, the psychomotor ability has enjoyed vast research efforts (Fleishman, 1956) and has been recognized as a critical ability for pilot performance (Griffin & Koonce, 1996; Wheeler & Ree, 1997). Most pilot performance-based selection batteries include tests targeting this ability. This category of TBs is best represented by USAF batteries: the Basic Attributes Test (BAT; Carretta, Zelenski, & Ree, 2000) and its successor the Test of Basic Aviation Skills (TBAS; Carretta, 2005), the U.S. Navy's performance-based ASTB (Phillips et al., 2011), as well as the Pilot Aptitude (PILAPT; Kokorian, Valsler, & Burke, 2016) and MICROPAT (Bartram, 1995), which was designed by a British psychologist. Given the importance of psychomotor ability and the numerous validation studies available supporting its predictivity for pilot performance, it was expected that this predictor group would show significant positive correlations with pilot performance criteria.

2.7.4 TBs Saturated with Controlled Attention

The content of the broad domain of Controlled Attention includes fluid reasoning, working memory, and processing speed. Findings suggest that processing speed and working memory capacity are essential precursors to fluid reasoning (Schneider et al., 2016). Additionally, tests of these abilities have long been associated with multitasking capability (Colom, Martínez-Molina, Shih, & Santacreu, 2010), which is recognized as an important ability for flying (Barron & Rose, 2017). There is strong evidence that working memory and fluid reasoning are crucial to the ability

to perform simultaneous dual tasks (Konig, Buhner, & Murling, 2005). Due to the established link between multitasking ability and the broad abilities forming the factor of Controlled Attention, this study limited the inclusion of TBs in this category group exclusively to those involving some forms of multitasking (e.g., dual cognitive tasks, dual psychomotor tasks). Many TBs include tests for assessing this ability, including the WOMBAT (O'Hare, 1997), MICROPAT (Bartram, 1995), BAT (Carretta et al., 2000), TBAS (Carretta, 2005), and the U.S. Navy's Computer-Based Performance Test (Delaney, 1992). On the basis of previous validation studies' results, it was expected that TBs saturated with Controlled Attention (i.e., multitasking) would be a valid predictor for pilot performance.

2.7.5 TBs Saturated with General Ability

General mental intelligence (*g*) is the third stratum in the CHC model, which sits at the top of all stratum I narrow abilities and stratum II broad abilities (McGrew, 2009). Although the label General Ability for this category may imply intelligence (*g*) tests, the TBs included here are not necessarily characterized as measures of intelligence. Some test batteries do provide an intelligence quotient summary (e.g., MAB-II, MicroCog), but others do not. Data may even include TBs with a score not entirely related to ability domains, such as the Pilot Candidate Selection Method (PCSM), which includes a rating for the previous flying experience. So, this category is best seen from an integration point of view as many TBs under this category were jointly computed for the purpose of integrating scores of different broad abilities (stratum II) and providing an overall composite that may increase validity. The type of tests in this group was mixed and hybrid with no specific orientation other than General Ability. Examples of TBs defined as saturated with General Ability are the PCSM (Carretta, 2011), MAB-II (full-scale IQ composite; Carretta et al., 2016), and the MicroCog (King et al., 2013). Since this group of tests contains a

variety of ability tests containing some indicators for *g*, it was expected that it would show criterion-related validity of positive magnitude.

2.7.6 TBs Saturated with Work Sample

In aviation, the best representation of a work sample is the flight simulator, as it provides an environment similar to the actual flying environment, especially with advancements in technology. Tirre (1997) differentiated between the "learning sample" (i.e., simulator) and "basic attributes" as two distinct approaches for aviator selection. He highlighted the advantage of the learning sample as a dynamic measurement of cognitive processing skills. This category differs from the other types that were linked to the broad domain of abilities in the CHC model. Instead, it was more grounded in the literature on work sample and job analysis methods (e.g., Roth et al., 2005; Schmidt, Hunter, & Outerbridge, 1986). The work sample in this group was exclusively given to the advanced flight simulator and does not include any other forms of actual flying (e.g., pre-training familiarization program). In terms of test batteries, the flight simulator is often considered as a controlled testing technique whereby the applicant is required to perform tasks simulating the job for which he or she is applying (Woychesin, 2002). The applicant's performance is tested in order to estimate the likelihood of success as a pilot candidate. The best available example of a simulation-based approach to pilot selection is the Canadian Automated Pilot Selection System (CAPSS), which is a computerized moving base simulator of a single-engine light aircraft (Spinner, 1991). It consists of four 1-hour sessions that progress in complexity. Carretta and Ree (2003) showed that the predictive validity of flight simulators is as high as that of general cognitive ability, and it is comparable to the validity of work sample tests found in other professions. Accordingly, it was anticipated that TBs of Work Sample methods (i.e., flight

simulator) would be a valid predictor for pilot performance, with a magnitude higher than other TBs.

2.8 Further Consideration of the Classification Schema

It is important, however, to reemphasize that the proposed category schema does not necessitate that all tests within a given TB have to be purely measuring the ability domain to which they are ascribed. Instead, it requires that the TB is dominated by tests indicative of that higher-order ability domain with the possibility of having fewer tests from other domains or even beyond, such as biodata or personality measures. For instance, a former version of the U.S. Navy/Marines selection test battery (ASTB) added the biographical inventory score to the composite pilot aptitude rating (PFAR) used in pilot selection (Stricker, 2005). Also, the Pilot Candidate Selection Method (PCSM) used by the USAF adds a measure of prior flying experience to the combination of two composite scores. Hence, the saturation approach is best viewed as a functional grouping consisting of, for the most part, two clusters of abilities within each TB: the majority cluster, which represents the tests highly saturated with the corresponding broad ability construct, and the minority cluster, which represents tests that may diverge from this line of abilities.

Relatedly, a well-accepted psychometric hypothesis maintains that an intelligence index may be sourced from TB that includes a sufficient number of mixed types of cognitive ability tests (Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, Nijenhuis, & Bouchard, 2008; Ree & Earles, 1991; Thorndike, 1987; Vernon, 1989). This is seen as evidence for the existence of a unitary higher-level general intelligence construct whose measurement is not dependent on the specific abilities assessed (Johnson et al., 2008). According to this hypothesis, many TBs analyzed here are qualified to be intelligence indices. However, this study does not claim that any of its proposed categories represents intelligence, even though several validity

coefficients included in the General Ability category are used as intelligence quotients. The strict definition of intelligence may not be appropriate for the more flexible and practical framework approached in this investigation.

Another well-established hypothesis is that cognitive abilities intercorrelate positively. Such positive relationships between abilities is explained by the existence of an underlying general intelligence factor (Jensen, 1998; Van Der Maas et al., 2006; Van der Maas, Kan, & Borsboom, 2014). For that reason, many theorists currently advocate a much broader, more multidimensional conception of intelligence (Schneider & Newman, 2015). Reeve et al. (2015) highlighted the need to return to more specific aspects of the intelligence network that allows for the development of assessments that are both theoretically rigorous and relevant to the complex manifestations of behavior that practitioners need to understand and predict. In view of this, the suggested categories may be seen as an attempt to capture some dimensions of the intelligence network relevant to the complex manifestations of pilot behavior, and also correspond to the need for modeling abilities predictive of future pilot performance.

2.9 Pilot Performance Criteria

Traditionally, the most-used criterion for evaluating pilot selection tests has been dichotomous training success (pass/fail, graduate/eliminate, completed/not completed). Studies have shown that the criterion chosen make a difference in understanding the relationship between selection factors and pilot training performance (Carretta, 1992; Weissmuller & Damos, 2014). Carretta (1992) examined several alternate performance criteria based on flying performance data (i.e., daily flying, flight check, and academic performance). Results show that different performance criteria yield different validity coefficients. Reviews of validation studies of pilot

selection tests also indicate that the effacacy of performance criteria varys and different magnitude of validity is obtained from different criteria (e.g., Carretta & Ree, 1995; Gibb & Dolgin, 1989; Johnston & Catano, 2013; King et al., 2013; Teachout et al., 2013). For that reason, this study included the investigation of four specific criteria that were found to be the most examined and reported in validation studies. Specifically, the criterion of (a) flying performance rating, (b) graduate/attrite training, (c) academic performance grade, and (d) flight simulator performance rating were investigated.

The four criteria chosen for this study are the most frequently used measures for pilot performance. The flying performance rating is arguably the best measure of flight performance as it combines pilot's performance across the daily flying and regular flight checks throughout the training program. Successful graduation from training program or failing to graduate for one reason or another is another important measure of flight performance. It provides a reasonable index for the overall performance of pilot that include flying, academic, simulator, and mental attitude. Academic courses at ground schools are essential part of training that teach trainee theories of flying, aircraft, and related sciences. Although relying solely on this criterion for measuring pilot performance is not sufficient, it does help in giving a partial indicator for performance that can be combined with other indexes. A flight simulator provides the simulation of actual flying training and gives trainee hands-on experience that imitates real-world operations. Despite the expected differences between the two experiences (i.e., actual and simulated), this method is useful for assessing flight performance.

In order to overcome the limitation of small sample size for some specific criteria and to make use of all collected studies, an overall criterion was constructed representing the best index of flying performance in each study. Among the performance criteria mentioned above, the most

indicative ones for actual flying performance are flying performance rating and graduate/attrite training. Flying performance rating was given priority for selection than graduate/attrite training due to its favorable level of measurement. This procedure was preferred over the common procedure of averaging criteria from the same sample. It was thought that the effect sizes produced by this procedure might be more indicative of actual flying performance and produce a more valid indication of the association between the test batteries and pilot actual flying performance. For example, utilizing the effect size derived from actual flying performance rating might give a better indication of pilot performance than a combined effect size computed from averaging flying performance rating, flight simulator performance rating, and academic performance. In general, the expectation was that the meta-analyzed mean validity would vary as a function of criteria types.

2.10 Development of Moderator Variables

Five variables were identified as potential moderators that may impact the associations between predictor variables and criterion variables. Hypotheses about potential moderating variables are usually formulated to detect any differences in validities attributed to specific group characteristics or contextual aspects (e.g., Aguinis, Gottfredson, & Wright, 2011). To establish the effect of the moderator, the data are sub-grouped and then reanalyzed separately according to the moderator variable (Schmidt & Hunter, 2015). Some assumptions were derived from previously related meta-analyses, while others were informed by the general literature on the topic. Following is a brief description of each moderator variable.

2.10.1 Number of Tests Incorporated in the Battery

The primary objective of including multiple tests in the assessment battery is to be able to capture a wide-ranging picture of the individuals' abilities and traits. Ideally, increasing the number of tests contributes positively to the overall assessments and supports the psychometric properties of the battery. Adding more tests to the battery is a common technique to improve the multiple correlations and the prediction strength of the battery (Horst, 1951a; 1951b). However, findings in this area are inconsistent. Based on the distribution of the number of tests found in the aggregated studies, TBs were coded as either a large test battery (five tests or more) or small test battery (fewer than five tests). Cut off point of five tests might be a reasonable point for distinguishing between large and small TBs. It was hypothesized that large TBs would predict pilot performance better than small TBs.

2.10.2 Regularity of TB Use in Pilot Selection

It is fairly reasonable to expect that TBs initially designed for pilot selection would predict pilot performance better than TBs intended for purposes other than pilot selection. Martinussen (1996) noticed differences in validity trends over time between tests used in pilot selection and those not used directly in selection. Although a large number of primary studies collected for this study mostly examined TBs used in a selection context, there were a fair number of studies involving TBs used for purposes other than selection (screening, experimental). As such, TBs were coded as being commonly used or uncommonly used in pilot selection, and it was hypothesized that TBs commonly used in pilot selection would predict pilot performance better than TBs that were uncommonly used in pilot selection.

2.10.3 Year of Publication

This variable was tested by Hunter and Burke (1994) and Martinussen (1996), who noted that validities tended to decline over time. They indicated, however, that the result may not imply an actual decline in predictive validity. Rather, it may be an effect of possible changes in pilots' selection procedures over the years or a deflation caused by a limited variability in the applicant population in later studies. Validity decline may also be attributable to potential differences in the population of applicants such as educational requirements and previous experience or differences in the flying training operational environment such as the criterion predicted (Hunter & Burke, 1994). The decade of study had the most significant impact among four examined moderators in Hunter and Burke (1994). In an attempt to understand this trend, they observed a general decline in validity since 1961 in five of the predictor groups. They also noted that the research before 1961 were dominated by research during World War 11 and had much larger average sample sizes while the research in more recent decades were characterized by smaller study samples and thus, larger sampling error that may have increased the variability of validity estimates. In order to investigate a potential change in the TBs' validities over time, the primary studies were separated into three subgroups, roughly by decade: 1987-1999, 2000-2009, and 2010-2017. It was hypothesized that TB validity estimates would tend to decrease over the years of publication.

2.10.4 Flying Organization

Hunter and Burke (1994) compared two subgroups of services (Air Force and other) on only three predictors. Martinussen (1996) compared three subgroups of military services (Air Force, Navy, and Army) on nine predictors and noted some differences in validity. Based on the collected data, this study allowed testing for four distinct flying organization groups: U.S. Air

Force, U.S. Navy, Another military, and Civilian. Since there was no reason to expect a specific trend for validity estimates among the subgroups, no direct hypothesis was posited for this moderator. Instead, a research question was proposed for this investigation.

2.10.5 Criterion level of measurement

The most frequent criterion in flying-related validation studies is training success, which usually takes a dichotomous format (e.g., graduate/attrite, complete/not complete, pass/fail). From a statistical point of view, dichotomization of variables leads to a reduction in the correlation such that even the 50–50 split leads to an approximately 20% reduction in the correlation (Hunter & Schmidt, 2004, p. 36). The second most common criterion usually takes the form of a continuous grade or rating for either actual flying performance or academic performance. Ordinal criterion has also been attempted by some studies, particularly for those using class rank as an index for performance. A less frequent criterion is the contingency table, which may be a favorable option in some cases. For instance, two groups of applicants may be formed according to their performance on a given test (e.g., 70th percentile) and subsequent comparison with a performance index such as the number of mishaps (e.g., less than or greater than five times). Due to the mixed results regarding the effect of the criterion level of measurement, there is not a direct hypothesis was suggested for this variable.

CHAPTER 3. METHOD

Three important steps were carried out to prepare for the present meta-analysis: (a) conducting an exhaustive literature search for the criterion-related validity studies of test batteries (TBs) used in aviation, (b) extracting and coding information from these studies, and (c) analyzing and summarizing the findings. A brief summary of the study's procedures is offered below.

3.1 Literature Search

A literature search was conducted to identify published and unpublished criterion-related validity studies that used TBs for pilot selection or assessment. Due to the existence of two meta-analyses that had synthesized results of older publications (Hunter & Burke, 1994; Martinussen, 1996), the search was limited to the period from 1987 to the present. First, an extensive search was conducted to locate published and unpublished research. Several search methods were utilized, including electronic searches and manual searches. For electronic database searches, a systematic search was conducted for publications in Google Scholar, the Defense Technical Information Center, PsycINFO, ProQuest, and Google search engine. The keywords used, both individually and in combination, were "pilot selection," "pilot assessment," "selection tests," "selection battery," "test battery," "flight aptitude test," "flight training program," "flight screening program," and "pilot performance." The well-known test batteries used in pilot selection and assessments were also used as keywords ("AFOQT," "ASTB," "MICROPAT," "CogScreen," "MicroCog," "MAB," and "PILAPT").

Publisher-specific databases (e.g., EBSCOhost, JSTOR, Web of Science, Elsevier Science Direct, and Wiley Interscience) were searched to identify any other sources. For unpublished dissertation and theses, a search was conducted using specialized databases such as ProQuest

Dissertations & Thesis and Theses Canada. For manual searches, the reference lists of the studies retrieved were reviewed for additional relevant studies. Furthermore, a manual search was carried out of the key journals in aviation, including *The International Journal of Aviation Psychology*, *Aviation Psychology and Applied Human Factors, International Journal of Applied Aviation Studies*, and *Military Psychology*. The annual conference abstracts presented to *the International Military Testing Association* were manually searched through the website. Moreover, the studies included in the Howse and Damos (2011) bibliographic database for the history of pilot training selection were inspected for any further resources.

3.2 Inclusion Criteria

The abstracts of studies were reviewed; studies that examined the criterion-related validity of pilot selection and assessment were considered eligible for inclusion. The objective was to identify primary studies whose results were relevant to the validity of TBs for predicting pilot performance and whose assessment practices complied with professional standards for test validation. Hence, several criteria were specified to direct the review phase of the collected studies. Primary studies were considered for inclusion if they fulfilled the following criteria:

(1) Independent variable was a composite score of *at least* two tests or two tasks. Primary studies that reported only validities from individual tests were excluded;

(2) Dependent variable was indicative of pilot performance, whether actual or simulated;

(3) Sufficient information about predictor measure and performance outcome was provided;

(4) Primary study reported a univariate effect between the independent and dependent variables or provided statistics that allowed the calculation of correlations (*t* test, *F* test, $\chi 2$ values); and

(5) Sample size was reported.

3.3 Coding Procedure

Using the search mentioned above and inclusion criteria, the search yielded 170 research reports, of which 89 were acceptable for inclusion in the present study. The primary studies included in the analysis are summarized in Appendix A grouped by TB type. Some studies involved more than one independent sample, each of which was treated separately. Duplication and even triplication of studies was identified (e.g., published paper, technical report, conference paper). Many published studies conducted with the U.S. Air Force were found to be preceded by technical reports or conference papers. In some cases, this was advantageous, since technical reports occasionally provide more information than the published papers. The most complete paper was chosen for coding purposes. For example, the published paper of Carretta et al. (2014) was preceded by a comprehensive technical report (Teachout et al., 2013). Also, King et al.'s (2013) published paper was preceded by a thorough technical report (King, 2012). The basic information from the primary studies was first coded (e.g., bibliographic information, sample characteristics). Additional information related to whether the sample comprised student pilots (applicants) or experienced pilots (incumbents), as well as whether the TB and criterion data were collected using a concurrent or predictive design was also included. Potential moderators (five variables) were carefully reviewed and coded. Determining the construct saturation of each TB was an important coding decision. TBs were categorized only when the primary study included adequate information about the tests contained within the battery or when it was clear that TBs manifested a certain orientation according to the descriptions provided. In some cases, there was a need for additional review of the batteries' manuals, websites, or other related articles in order to confidently assign them to the correct category. For instance, the CogScreen test battery, a widely used selection TB

for pilots in civilian organizations, required an additional search to understand the constructs and factor structure underlying the nearly 65 scores derived from it (e.g., King et al., 2012).

In some primary studies (e.g., Carretta, 2005; Ingurgio & Crawford, 2017; Phillips et al., 2011), different test combinations were attempted and reported. I chose to select the composite commonly used in pilot selection or assessment. When there was no known common composite due to the exploratory nature of the study, I selected the composite that was recommended by the authors. If it happened that a composite included non-cognitive measure, I chose the composite that was less contaminated with measures other than cognitive tests whenever possible. For example, Carretta (2011) reported the correlations between flying performance and PCSM composite scores with and without personality constructs. Although the correlations involved personality constructs were slightly higher, I selected the coefficient without personality constructs. This procedure should constrain the scope of the study and limit the examined variables to those in the cognitive domain. For university flight programs, there is usually no special battery for the selection of students aside from the regular requirements for college entry. To obtain useful information for this study (e.g., Forsman, 2012; McFarland, 2017), I added the composite score of standardized tests (e.g., ACT, SAT) to TBs of Acquired Knowledge whenever reported to serve as alternatives to the selection battery found in military and airline settings.

The moderator of the number of tests in the TBs (small TB/large TB) was not expected to be present in two types of TBs, those saturated with Controlled Attention and General Ability. For TBs of Controlled Attention (dual tasking), it is clear that handling multiple tasks simultaneously is only possible to a certain extent. Thus, I did not expect to find many TBs for the large battery (5 tests or more) subgroup. Similarly, it is typical for the TBs of General Ability to contain a large number of tests in order to increase measurement precision. Hence, it was not expected to see

many TBs in this category for the small battery (less than 5) subgroup. Regarding the moderating effect of the regularity of TB use in pilot selection, it was important to understand the context of the study and the extent of using the composite scores in pilot selection. Occasionally, selection/assessment TBs give multiple composite scores, each of which are used for different purposes and decisions. Composite scores that are commonly used in pilot selection were marked differently from those that are used less often. For example, the AFOQT provides multiple composites, including a pilot composite for pilot selection and a navigator/technical composite for navigator and technician selection. If it happened that a study reported both composites (e.g., Carretta, 1988; 1992; 1997; Keener, 2003), then the pilot composite was considered to be the commonly used battery in pilot selection, and the navigator composite was seen as the uncommonly used battery in pilot selection.

The study considered four specific types of performance criteria and one global index for pilot performance. For the overall index, one single criterion variable was selected from each study that was thought to represent the best index for pilot performance presented by that study. Forming this global criterion was necessary in order to increase the sample of studies in the meta-analyses and to facilitate further analyses of potential moderators. The fairly small number of studies aggregated for each specific criterion would not allow practical testing for the moderators. The selection of the one 'best' criterion from each study followed this order of preference: (1) flying performance rating, (2) graduate/attrite training, (3) flight simulator performance rating, and (4) academic performance grade. For the flying performance criterion, preference was given respectively to (1) continuous-scaled criterion, (2) ordinal-scaled criterion, (3) dichotomous-scaled criterion, and (4) contingency table.

Finally, the recommendation in meta-analysis is to have the primary studies coded by multiple coders to estimate their level of agreement (Aytug, Rothstein, Zhou, & Kern, 2012). A common rule of thumb requires 80% level of agreement between the coders or even higher (Bayerl & Paul 2011). In the present study, the complete list of collected studies was coded by the author and 35% (31 studies) randomly selected subsample were recoded by a student who is in his third year in Ph.D. program in the College of Education. The agreement between the two coding was adequate, ranging between 89% to 100% across the coded variables. The agreement was 100% for the year of publication, 96% for sample size, and 93% for validity coefficients. The least agreement existed for categorization of TBs (89%) and the regularity of TB use in pilot selection (91%). The outcome of this appraisal was indicative of a satisfactory level of accuracy that allowed for the meta-analysis. The consensus concerning disagreements was sorted out by discussion.

3.4 Effect Size

The collected effect sizes from the primary studies were coefficients of validity computed as a correlation between composite scores derived from multiple test batteries and a measure of pilot performance. The predictors consisted of six types of TBs with different construct saturations: (a) Acquired Knowledge, (b) Perceptual Processing, (c) Motor Abilities, (d) Controlled Attention, (e) General Ability, and (f) Work Sample. Outcome measures consisted of four specific and one overall type of criteria: (a) flying performance rating, (b) graduate/attrite training, (c) flight simulator performance rating, and (d) academic performance grade (e) overall criterion for pilot performance. As the direction of the rating format of both tests and criteria were not always consistent, the sign of correlation coefficients was reverse coded in some cases so as to obtain a homogenous interpretation of results for TB-outcome associations.

3.5 Analysis Plan

The first series of meta-analyses were planned to assess the criterion-related validity of each category of TBs for four criteria of pilot performance. A total of 24 meta-analyses would have been conducted if the collected data were sufficient for each combination. However, some meta-analyses could not be performed due to the absence of studies thus, reducing the analyses to 21 relations. The secondary meta-analyses were broader in scope in that they relied on one single criterion selected from every primary study collected for each type of TB. A total of 6 meta-analyses were planned, one for each category of TB saturation. The third series of meta-analyses tested the effect of moderators. Based on each moderator, subgroups were formed, each of which underwent a separate meta-analysis. If the data allowed analysis for all subgroups, a total of (96) meta-analyses would have been conducted. However, subgroup analysis could not be completed on a number of cases due to insufficient sample size (≤1) or inapplicability. Hence, the moderator-based meta-analyses were possible for 74 out of 96 TBs-criteria relationship.

3.6 Meta-Analysis Procedure

The study applied the psychometric meta-analysis approach of Schmidt and Hunter (2015). This approach assumes random-effect models, which allow for the true effect size to differ across studies (Hedges, 1983) and take into account the true differences among studies and participants (Schmidt, Oh, & Hayes, 2009). This model considers additional sources of variance between studies (Viechtbauer, 2005) and is seen as more appropriate than a fixed-effects model for investigations involving high levels of heterogeneity. Given the multiplicity and diversity of the tests included in this study, heterogeneity due to systematic differences among studies was expected; thus, the random-effects model is a sensible choice to consider the various sources of variance between studies. Schmidt et al. (2009) found that meta-analyses based on the random-

effects model tend to produce more accurate and less biased estimates than meta-analyses based on the fixed-effects model. The main feature of the Schmidt and Hunter approach is that it allows for correcting for studies' artifacts (e.g., sampling error, predictor unreliability, criterion unreliability, range restriction) to estimate population correlations (ρ; Schmidt, 2015). However, sample specific data were insufficient to correct studies for either range restriction or unreliability either individually or through artifact distributions. Hence, no correction for attenuation was made to the mean validity other than that due to sampling error (i.e., "bare-bones" meta-analysis). The estimates, therefore, are likely to be underestimates of the true theoretical relationship between composite scores of TBs and pilot performance criteria.

In spite of this, the sampling error is considered the most critical artifact that can impact the outcomes of a meta-analysis, if not properly controlled. Of the total artifactual variance, Koslowsky and Sagie (1994) found that sampling error alone accounted for more than 90% for small or medium samples and more than 70% for large samples. Although Schmidt and Hunter consider this form of meta-analysis as lacking, other researchers have shown that corrections for artifacts could be imprecise with a small number of studies analyzed (Spector & Levine, 1987), which was the case for some subset analyses in the present meta-analysis. Data were analyzed for all relationships that had been measured in at least two independent samples. Also, given the nature of the current study and the relationships estimated, a conservative approach may be more justifiable. If the test batteries are found to perform adequately, further work with additional artifact corrections would seem in order. The focus on only one artifact will draw more attention to the relative importance of each predictor for flight performance and whether they are significant predictors, even with underestimated correlations. Furthermore, it would seem unrealistic to see that test battery or flying performance measures are perfectly reliable and hence, an estimation of

validities in an ideal world in which no statistical artifacts remain is less practically useful than an understanding of validities in the observed world. Selection decisions of pilots are made on observed scale scores, rather than the theoretical standings of participants on the constructs measured by the TBs. Lastly, relying on observed correlations should be more consistent with previous meta-analyses in pilot selection tests and more comparable to Hedges's random-effect approach (Hedges & Olkin, 1985; Hedges & Vevea, 1998). Accordingly, all validity estimates (the estimated true score correlation) were assessed using the observed (uncorrected) correlation coefficients. A "bare-bones" meta-analytic procedure was used to estimate the mean validity, the observed variability around the mean, and the variability left over after accounting for variability due to sampling error.

According to Schmidt and Hunter (2015), a population correlation between predictor and criterion variables is best estimated using the following model:

$$\bar{r} = \frac{\sum[N_i r_i]}{\sum N_i}$$

where $\bar{r}$ is the arithmetic mean of all predictor-criterion correlations, r_i is the correlation in study i and N_i is the number of persons in study i. Similarly, variance (s^2) across studies (significance of mean effect size) is best estimated using frequency-weighted average squared error, as seen below:

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

Variability around the mean correlation is often expressed using confidence intervals and credibility intervals (values). Confidence intervals estimate the variability in the mean correlation due to sampling error, while credibility values estimate the variability in the individual population correlations across studies independent of sampling error (Whitener, 1990). These are the estimates of the range of the distribution of the true effect size ($\rho \pm 1.28$ SD_ρ), which is interpreted

as 80% of the values in the ρ distribution lying between the lower and upper band of this interval (Schmidt & Hunter, 2015). The importance of credibility interval (i.e., prediction intervals) over confidence intervals in meta-analysis is frequently emphasized (e.g., Schmidt & Hunter, 2015; Koslowsky & Sagie, 1993; Whitener, 1990) as it reflects the underlying population effect sizes and contains a percentage of the distribution of a random variable.

3.7 Evaluation of the Results

Meta-analysis results were assessed based on (a) the magnitude of mean validity estimates, (b) the 95% confidence intervals (CI), (c) the 80% credibility value (CV), and (d) the percentage of variance explained according to the 75% rule. Regarding the mean correlations, effect sizes suggested by Lipsey and Wilson (2001) were referred to, namely, .10 is small/low, .25 is medium/moderate, and .40 is a large/high effect. For the significance of mean correlations, the 95% CIs were assessed to determine whether they included zero. A non-zero 95% CI indicates significant mean correlations. To evaluate the variability of the mean validities across samples (i.e., heterogeneity), 80% CVs were examined to see whether they included zero, as a non-zero 80% CV indicates that validity may be generalizable across settings and samples. If CI does not contain zero, but CV includes zero, that gives an indication of valid effect size but with limited validity generalization. Heterogeneity in the mean correlations (ρ) was further assessed using the 75% rule suggested by Schmidt and Hunter (2015). According to this rule, a search for moderators is only warranted if less than 75% of the variance is explained by artifacts. If this value exceeds 75%, the remaining unexplained variance is most likely due to uncorrected artifacts in the studies (including sampling error) and should be ignored. To test the degree of moderating effects, the 95% CI around the mean validity (ρ) of each group was used (Whitener, 1990). The noticeable mean difference and nonoverlapping of confidence intervals were regarded as indicators for

moderating effects and non-artifactual difference between the compared true validities. In contrast, overlapping CIs were seen as an indicator of an insignificant moderating effect.

3.8 More about Meta-analysis

The meta-analysis investigations of organizational sciences predominantly use the psychometric meta-analysis methods of Schmidt and Hunter. In Aytug et al.'s. (2012) review of meta-analyses practices in organizational research, they found nearly 81% of the meta-analyses used Hunter and Schmidt's methods exclusively, 16% used Hedges and Olkin's methods, and a few meta-analyses (2%) used both methods. Schmidt and Hunter (2015) method tends to work best for the Pearson-*r* correlation-based meta-analysis, whereas the Hedges and Vevea (1998) method tends to work best for *d* effect size-based meta-analysis (Brannick, Yang, & Cafri, 2011). One essential strength of the Schmidt and Hunter (2015) method is its assumption of random-effects models, which have become increasingly popular and are considered the most effective approach in reaching the scientific aims of meta-analyses (Erez, Bloom, & Wells, 1996; Hedges & Vevea, 1998; Hunter & Schmidt, 2000). In comparison between the two widely used random effects models, results indicate that Schmidt and Hunter's (2015) approach has generally provided more accurate results than has Hedges and Vevea's (1998) approach (Field, 2001; 2005; Hall & Brannick, 2002). Hafdahl and Williams (2009) did replicate Field's (2001) simulation results under homogeneous correlation parameters, but those under heterogeneity were found not replicable. They showed that a more appropriate *z*-to-*r* transformation can improve the modest performance of Hedges and Vevea's (1998).

Despite the several strengths of Schmidt and Hunter approach, there has been criticism to the method for using untransformed Pearson-*r* correlations for estimating the average effect size of the aggregated effect sizes, which likely to skew the distribution and give biased estimates.

However, there is a controversy about whether it is better to use untransformed *r* or *r* transformed to Fisher's *z*. According to Hunter and Schmidt (2004), the use of the Fisher *z* transformation also leads to a substantial bias the mean correlation (i.e., overestimation) and may cause serious inaccuracies in random-effects meta-analysis models (p. 56). They claim that meta-analysis is never made more accurate by using the Fisher *z* transformation (p. 83). Several simulation studies have also supported this conclusion (Field, 2001; 2005; Hall & Brannick, 2002). Moreover, the method of Schmidt and Hunter (2015) has been criticized for its weighting choice for computing the average effect size and the variance underlying the effect sizes. Contrary to the methods using inverse variance weights (e.g., Hedges & Vevea, 1998) or unit weights (e.g., Bonett, 2008), Schmidt and Hunter (2015) method uses sample size weights for estimating the overall mean, which assigns a greater weight for studies that provide larger sample size. Brannick et al. (2011) argued, however, that if we have many large-sample studies, weights become less important and any sound weighting scheme will yield the same mean value. They also asserted that it is still unknown whether weighting schemes applied to effect sizes in common meta-analyses really matter and make any practical difference.

3.9 Supplementary Analysis

In order to compare different approaches of meta-analysis and to give the concluded results further support, the random-effects model of the Hedges and Olkin (1985) approach to meta-analysis was applied to the overall criterion data. As in the Schmidt and Hunter approach, this approach provides an estimate of both the overall mean effect size and the variability of infinite-sample effect sizes (Brannick et al., 2011). For the correlation-based meta-analysis, it first transforms correlations using Fisher's *z* transformation; the calculations are carried out in Fisher's *z* metric, and then back-transformed to the *r* metric (Hedges & Olkin, 1985). In their comparisons

of three meta-analysis approaches (Hedges & Olkin, Rosenthal & Rubin, and Schmidt & Hunter), Johnson, Mullen, and Salas (1995) found that the techniques of Hedges and Olkin and Rosenthal and Rubin were quite similar to each other, while the Schmidt and Hunter method diverged from the others to a noticeable extent. Specifically, they realized that the Schmidt and Hunter approach tended to yield more conservative estimates of the significance of effect sizes and wide variant estimates of moderators.

For the Hedges and Olkin approach applied in this study, the restricted maximum-likelihood estimator (Viechtbauer, 2005) was used to compute the between-study variance. Heterogeneity was assessed using the *Q*-within test (for statistical significance) and the I^2 index (for practical significance; Higgins & Thompson, 2002). If the *Q* statistic is statistically significant and if I^2 is more than 75%, moderators are likely to be present (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

3.10 Publication Bias

Additionally, in the same context of the Hedges and Olkin's (1985) approach of meta-analysis, publication bias was investigated to assess the sensitivity of the findings. Publication bias has an effect when the meta-analyzed studies depart systematically from the typical research on a particular topic. Although the study sample included all relevant primary studies that were located, regardless of their publication status, four indices of publication bias were examined: funnel plots, Begg rank correlation test, Egger weighted linear regression test and fail-safe analysis. The visual presentation of the funnel plots was inspected for patterns of asymmetry consistent with publication bias (Sterne, Becker, & Egger, 2005). To test funnel plots' asymmetry statistically, a Begg rank correlation test (Begg & Mazumdar, 1994) and Egger weighted linear regression test

(Egger, Smith, Schneider, & Minder, 1997) were used to judge potential small-study effects and publication bias across studies. Further, the results of the fail-safe analysis were examined to determine the additional sample size with zero correlation between TB and criterion scores that would have to exist to produce a mean effect size that was not statistically significant (p = .05; Orwin, 1983). It indicates the number of missing sample members showing no relationship between the predictor and criterion scores that would need to exist to nullify the observed effect.

3.11 Software

For the Schmidt-Hunter approach, meta-analyses was performed using a program developed specifically for this method (Schmidt & Le, 2004). For the Hedges and Olkin approach, the supplemental meta-analysis was performed using the Jamovi program, which is based on the R metafor package (Viechtbauer, 2010).

CHAPTER 4. RESULTS

This chapter presents the results corresponding to each hypothesis and research question on the predictor–criterion–moderator relationships. Table 1 summarizes the variables examined in the study, which included six predictors, five criteria, and five moderators.

Table 1
Summary of variables investigated in the study

Predictors (TB category)	**Criterion (performance)**	**Moderators**
(1) Acquired Knowledge	(1) Flying performance rating	(1) Number of tests in the battery
(2) Perceptual Processing	(2) Graduate/attrite training	(2) Regularity of use in pilot selection
(3) Motor Abilities	(3) Academic performance grade	(3) Year of publication
(4) Controlled Attention	(4) Flight simulator performance	(4) Flying organization
(5) General Ability	(5) Overall flight performance	(5) Criterion level of measurement
(6) Work Sample		

Table 2 gives a summary of the hypothesis and questions investigated in the study, which included five hypotheses and four questions.

Table 2
Summary of the research hypothesis and questions

	Hypothesis/question statement
Hypothesis 1	The six ability saturations of TBs will show small to moderate mean correlations for predicting pilot performance across four specific criteria and they will generalize validity across samples and settings.
Hypothesis 2	The six ability saturations of TBs will show small to moderate mean correlations for predicting the overall pilot performance criterion and they will generalize validity across samples and settings.
Question 1	Among the six ability saturations of the test batteries, which is the best predictor for each performance criterion?
Question 2	Of the six ability saturations of the test batteries, which is the best predictor for the overall criterion of pilot performance?
Hypothesis 3	Large test batteries with 5 or more tests will predict pilot performance better than small test batteries with fewer than 5 tests.
Hypothesis 4	Test batteries commonly used in pilot selection will predict pilot performance better than test batteries uncommonly used in pilot selection.
Hypothesis 5	Validity estimates of the six test batteries will tend to decrease in more recent publications (1987-1999, 2000-2009, 2010-2017).
Question 3	Do the validities of test batteries vary as a function of the flying organization (USAF, US Navy, Another Military, Civilian)?
Question 4	Do validities of test batteries vary as a function of performance criterion level of measurement (Continuous, Ordinal, Dichotomous, Contingency table)?

4.1. Data Description

The meta-analysis included 116,806 participants analyzed in 89 studies that reported correlations between test battery composite scores (TBs) and pilot performance outcomes. These 89 studies included 41 journal articles, 27 military technical reports, 11 thesis/dissertations, 10 conference papers, and 1 conference poster. A total of 78 studies had single samples, 5 studies had two samples, 4 studies had 3 samples, one study had 4 samples, and one study had 14 samples yielding a total of 118 independent samples. From the 118 samples, 267 independent correlations were extracted. Of these, 138, 102, 27, and 18 were correlations with criteria of flying performance rating, graduate/attrite training, academic performance grade, and flight simulator performance rating, respectively. With regard to the settings where the research was conducted, 52 were from the USAF, 16 were from the US Navy, 7 were from the Canadian Force, 18 were from different militaries (i.e., 5 from the UK, 2 each from Norway, Italy, and Portugal; 1 each from Germany, Turkey, Poland, Korea, India, Chile, and Brazil), 19 were from civilian airlines (the USA, the UK, Germany, France, New Zealand, and Hong Kong), and 6 were from university flight programs. Validity coefficients (correlation) for some criteria (e.g., extra flying hours) were reverse coded if appropriate. Table 3 includes some details of the total number of respondents (*N*) and the number of independent samples (*k*) upon which each meta-analysis is based. More information about the accumulated primary studies can be found in Appendixes A.

4.2 TBs Relationship with Four Outcomes of Pilot Performance

First, the validity of each type of TB was meta-analyzed separately for four outcomes of pilot performance. Several samples appeared in more than one analysis as some studies reported multiple criteria. The results are reported in Table 3. Out of 24 planned analyses (six TBs X four criteria), three cases could not be analyzed due to the absence of any sample (Motor Abilities TBs

with academic grade, General Ability TBs *with* simulator rating, and Work Knowledge *with* simulator rating). The estimate of sample weighted-mean validities ranged from -.03 (Controlled Attention TBs *with* academic grade) to .42 (Motor Abilities TBs *with* simulator rating) across 2 to 47 independent samples covering total sample sizes between 232 and 60,835. The 95% CI showed that the true average validities of all TBs, but one exceeded zero, which support the significance of the mean validities, and may suggest an acceptable degree of predictivity. The relation of Controlled Attention TBs with academic grade criterion was the only exception to this overall result with mean validity of **-.03** and 95% CI of [-.11 to.04]. It is also apparent that only a small percentage of observed variation in validities is accounted for by sampling error, with only four validities satisfied the 75% rule (90% for Controlled Attention TBs *with* Simulator Raring, 79% for General Ability TBs *with* academic grade, 79% for Work Sample TBs *with* flying rating, and 100% for Work Sample TBs *with* academic grade). None of the remainders of the TBs satisfied the 75% rule. This suggested that there was variance left to be explained either by moderators or artifacts not corrected for.

Table 3
Meta-analyses of the criterion-related validity of TBs for four criteria of pilot performance

Type of Test Battery	*k*	*N*	ρ	SD_r	SD_ρ	95% CI	80% CV	%VE
Acquired Knowledge								
Flying Rating	47	48697	.14	.05	.043	[.12-.15]	[.08-.19]	33
Graduate/Attrite Training	33	60835	.12	.05	.048	[.10-.14]	[.06-.18]	18
Academic Grade	12	23935	.19	.06	.055	[.15-.22]	[.11-.26]	13
Simulator Rating	2	1634	.17	.17	.03	[.12-.23]	[.14-.21]	63
Perceptual Processing								
Flying Rating	30	26191	.14	.05	.042	[.12-.16]	[.09-.19]	39
Graduate/Attrite Training	16	46407	.11	.05	.044	[.09-.14]	[.06-.17]	15
Academic Grade	6	19525	.21	.02	.016	[.20-.23]	[.19-.23]	52
Simulator Rating	9	1514	.21	.18	.16	[.09-.33]	[.00-.42]	17
Motor Abilities								
Flying Rating	15	6282	.22	.09	.074	[.17-.26]	[.12-.31]	28
Graduate/Attrite Training	25	20965	.14	.09	.077	[.10-.17]	[.04-.23]	16
Academic Grade	0	-	-	-	-	-	-	-
Simulator Rating	3	1158	.42	.074	.06	[.34-.50]	[.34-.50]	33
Controlled Attention								
Flying Rating	16	11350	.09	.12	.11	[.03-.15]	[-.05-.23]	10
Graduate/Attrite Training	13	18026	.07	.05	.045	[.04-.10]	[.01-.13]	26
Academic Grade	2	7373	-.03	.06	.053	[-.11-.04]	[-.10-.03]	9
Simulator Rating	4	232	.34	.124	.04	[.22-.46]	[.29-.39]	90
General Ability								
Flying Rating	24	20830	.17	.06	.051	[.15-.20]	[.11-.24]	29
Graduate/Attrite Training	11	26724	.14	.08	.072	[.10-.19]	[.05-.23]	7
Academic Grade	4	14292	.23	.02	.008	[.21-.25]	[.22-.24]	79
Simulator Rating	0	-	-	-	-	-	-	-
Work Sample								
Flying Rating	6	1282	.35	.07	.031	[.29-.40]	[.31-.39]	79
Graduate/Attrite Training	4	871	.34	.25	.24	[.10-.59]	[.03-.65]	6
Academic Grade	3	635	.24	.02	0	[.21-.26]	[.24-.24]	100
Simulator Rating	0	-	-	-	-	-	-	-

Note. *k*=number of independent studies; *N*=total sample size; ρ= mean true-score correlation corrected only for sampling error; SD_r=sample-size-weighted observed standard deviation of correlations; SD_ρ =standard deviation of true-score correlations corrected for sampling error; CI=confidence interval around the mean true-score correlation; CV=80% credibility interval; VE= variation in the observed correlations attributable to sampling error.

The 80% credibility limits assist in evaluating how generalizable the TBs are outside of the study as it concerns other settings and samples. Of the 21 validities, only two associations related to Controlled Attention TBs included zero in the 80% credibility intervals specifically, with flying rating [-.05 to.23] and academic grade [-.10 to.03] which suggest that TBs of Controlled Attention do not effectively predict these two criteria. A comparison between criteria showed that

graduate/attrite training was the least well-predicted criterion across TB categories, except for Work Sample, which had an academic grade criterion as the least well-predicted criterion. Academic grade criterion along with simulator rating were the best-predicted criteria across each of three TBs. The mean validity magnitudes of Acquired Knowledge TBs were generally lower than the other type of TBs across the four criteria, although they were positive. Taken together, while TBs with different broad ability saturation were predictive of pilot performance across different outcomes, the strength of the relationship varied noticeably. Results suggested nongeneralizable validity for Controlled Attention TBs with flying rating and academic grade (CV included zero), and generalizable validity without further moderator analysis for the four relations that exceeded the 75% of explainable variance (Controlled Attention TBs *with* simulator rating, General Ability TBs *with* academic grade, Work Sample TBs *with* flying rating, and Work Sample TBs *with* academic grade). The remains 14 validities were likely to be *moderated*, and thus, the search for possible moderators is appropriate. Accordingly, Hypothesis 1 was only partially supported.

4.3 TBs Relationship with Global Index of Pilot Performance

Second, the validity of each category of TBs in relations with one single outcome of pilot performance selected for each independent sample (e.g., the best index presenting actual flying performance) was meta-analyzed. The results are reported in Table 4. Running the analyses by TB category demonstrated that sample weighted-mean correlations ranged from .10 (Controlled Attention TBs) to .34 (Work Sample TBs) across 9 to 68 independent samples covering between 1,655 to 93,209 participants. Consistent with the previous finding, the highest estimated mean validity among the six predictors was exhibited by Work Sample TBs (.34). For the rest of TBs,

the highest validities were obtained for TBs of Acquired Knowledge (.19) and General Ability (.18) whereas the lowest was obtained for TBs of Controlled Attention (.10) and Perceptual Processing (.14). According to the 95% CI, all validity estimates were significant. Some similarities were noticed in the 95% CI between TBs of Acquired Knowledge [.17-.21], General Ability [.15-.20], and Motor Abilities [.13-.20], which suggest that their respective mean validities were relatively comparable. Because the 95% CI for Controlled Attention TBs and Perceptual Processing TBs were lower and yielded little or no overlap with the 95% CI of other TBs, the results suggested that they were less predictive of pilot performance.

Table 4
Meta-analyses of the criterion-related validity of TBs for one overall index of Pilot Performance

Type of Test Battery	*k*	*N*	ρ	SD_r	SD_ρ	95% CI	80% CV	%VE
Acquired Knowledge	68	93209	.19	.09	.09	[.17-.21]	[.07-.30]	8
Perceptual Processing	47	48697	.14	.05	.04	[.12-.15]	[.08-.19]	33
Motor Abilities	39	24388	.17	.11	.10	[.13-.20]	[.03-.30]	12
Controlled Attention	29	20438	.10	.10	.09	[.06-.14]	[-.02-.22]	14
General Ability	31	34289	.18	.07	.06	[.15-.20]	[.10-.25]	20
Work Sample	9	1655	.34	.21	.19	[.21-.48]	[.09-.59]	10

Note. *k*=number of independent studies; *N*=total sample size; ρ= mean true-score correlation corrected only for sampling error; SD_r=sample-size-weighted observed standard deviation of correlations; SD_ρ=standard deviation of true-score correlations corrected for sampling error; CI=confidence interval around the mean true-score correlation; CV=80% credibility interval; VE= variation in the observed correlations attributable to sampling error.

None of the TBs categories satisfied the 75% rule, and hence, all met the criteria for conducting the moderator analysis. Only a small percentage of observed variation in validities was attributable to sampling error, the highest percentage being the 33% found for Perceptual Processing TBs. However, from the previous analysis, we knew that the explained variance of Work Sample TBs exceeded 75% for two of the three investigated outcomes. This indicated that most of the variability noted here on the global criteria was likely to be an effect of the third criteria (graduate/attrite training). The 80% CV for all TBs but one did not include zero, suggesting

generalized validity across samples. The exception was noted for Controlled Attention TBs with 80% CV of [-.02 to.22]. This part of analysis concludes that all TBs with different ability saturations significantly predict the global index of pilot performance. Because none of the TBs satisfied the 75% rule, this indicated that there was variance left to be explained either by moderators or artifacts that were not corrected for. Hence, Hypothesis 2 was only partially supported.

4.4 Moderator Analysis of TBs/Pilot Performance

The preceding analysis showed that sampling error associated with TBs' validities explained less than 75% of the variance in mean correlations across performance criteria. Thus, moderator analyses were carried out, attempting to understand the systematic variability in the mean validities. After subgrouping the data according to the moderator variables, series of analysis were performed on each TB-pilot performance relation using the global criteria of pilot performance. The investigation could not be conducted in a few instances due to inapplicability of a particular moderator to a certain predictor score, or inadequacy of the primary study sample.

4.4.1 Moderating Effect of Number of Tests in the TBs

This analysis examined whether the validity of TBs differed depending on the number of tests incorporated in the battery. TB was coded as being a "small" battery if it included two to four tests and a "large" battery if it included five tests or more. It was possible to test this moderator only for three TB categories (Acquired Knowledge, Perceptual Processing, and Motor Abilities). Hence, a total of six correlations were analyzed for this moderator (three TBs multiplied by two subgroups). As seen in Table 4, the three types of TBs correlated meaningfully with pilot

outcomes, regardless of whether the battery is small or large. The mean validity of small battery subgroups versus large battery subgroups were .25 and .15, respectively, for Acquired Knowledge TBs, .09 and .14 for Perceptual Processing TBs, and .26 and .12 for Motor Abilities TBs. There was an interesting trend in that small TBs appeared to show higher validity than large TBs in Acquired Knowledge and Motor Abilities. For Perceptual Processing, large TBs had higher weighted validity than small TBs, although the sample was relatively small (10 studies, N = 3,923). The 95% CI of both TBs' subgroups did not overlap in TBs of Acquired Knowledge and Motor Abilities, and barely overlapped in Perceptual Processing TBs ([.06-.12] for small vs. [.12-.16] for large). This indicated significant moderating effect for this variable. Overall, Hypothesis 3 was inadequately supported as only one category of TBs showed a higher validity estimate for the large battery than the small battery.

Table 5
Moderating effect of the number of tests incorporated in the battery (two subgroups)

Type of Test Battery	*k*	*N*	ρ	SD_r	SD_ρ	95% CI	80% CV	%VE
Acquired Knowledge	68	93209	**.19**	.09	.09	[.17-.21]	[.07-.30]	8
Small Battery (fewer than 5 tests)	28	33529	.25	.12	.12	[.21-.29]	[.10-.40]	5
Large Battery (5 tests or more)	40	59680	.15	.04	.04	[.14-.17]	[.11-.20]	34
Perceptual Processing	47	48697	.14	.05	.04	[.12-.15]	[.08-.19]	33
Small Battery (fewer than 5 tests)	10	3923	.09	.04	0	[.06-.12]	[.09-.09]	100
Large Battery (5 tests or more)	37	44774	.14	.05	.04	[.12-.16]	[.08-.20]	30
Motor Abilities	39	24388	.17	.11	.10	[.13-.20]	[.03-.30]	12
Small Battery (fewer than 5 tests)	20	7669	.26	.11	.10	[.21-.31]	[.14-.38]	100
Large Battery (5 tests or more)	19	16719	.12	.09	.08	[.09-.16]	[.02-.22]	15

Note. k=number of independent studies; N=total sample size; ρ= mean true-score correlation corrected only for sampling error; SD_r=sample-size-weighted observed standard deviation of correlations; SD_ρ=standard deviation of true-score correlations corrected for sampling error; CI=confidence interval around the mean true-score correlation; CV=80% credibility interval; VE= variation in the observed correlations attributable to sampling error.

4.4.2 Moderating Effect of the Regularity of TBs Use in Pilot Selection

This analysis examined whether TBs commonly used in pilot selection were associated with higher validity than those that are seldom used, or were unique to a particular primary study. TBs were coded as being commonly used if they were designed and administered for pilot selection and uncommonly used if they were designed for different purposes. It was possible to test this moderator for all TBs saturations with the exception of Work Sample TBs. Hence, a total of 10 validities was analyzed for this moderator (five TBs multiplied by two subgroups). The results are summarized in Table 6. TBs' subgroups (commonly used/uncommonly used) across the five ability saturations were valid predictors for pilot performance except for the uncommonly used subgroup of Controlled Attention TBs where both 95% CI and 80% CV included zero. Mean validities varied from .11 to .21 for the commonly-used subgroup and from .04 to .13 for the uncommonly-used subgroup. As may be expected, TBs that are commonly used in pilot selection were more predictive of pilot performance than TBs that are uncommonly used in pilot selection across the different ability saturations, except for Perceptual Processing TBs. Despite this result, the effect of this moderator was only significant in the case of Acquired Knowledge TBs according to the 95% CI. The rest of the TBs saturations showed overlapped intervals, which suggested insignificant differences between the mean of the two subgroups. This indicated that even TBs uncommonly used in the context of pilot selection might be useful predictors. Taken together, Hypothesis 4 was reasonably supported as most of the tested TBs saturation showed higher validities for the commonly-used subgroup than the uncommonly-used subgroup, although few differences were significant, perhaps because the subgroup sample sizes were small.

Table 6
Moderating effect of the regularity of TBs use in pilot selection (two subgroups)

Type of Test Battery	*k*	*N*	ρ	SD_r	SD_ρ	95% CI	80% CV	%VE
Acquired Knowledge	68	93209	.19	.09	.09	[.17-.21]	[.07-.30]	8
Commonly used	47	67291	.21	.10	.09	[.18-.24]	[.09-.33]	67
Uncommonly used	21	25918	.13	.04	.02	[.11-.14]	[.10-.15]	66
Perceptual Processing	47	48697	.14	.05	.04	[.12-.15]	[.08-.19]	33
Commonly used	9	3631	.11	.06	.04	[.07-.15]	[.06-.16]	60
Uncommonly used	38	45066	.14	.05	.04	[.12-.15]	[.08-.19]	100
Motor Abilities	39	24388	.17	.11	.10	[.13-.20]	[.03-.30]	12
Commonly used	34	23774	.17	.11	.11	[.13-.20]	[.03-.30]	11
Uncommonly used	5	614	.13	.11	.07	[.03-.23]	[.04-.22]	62
Controlled Attention	29	20438	.10	.10	.09	[.06-.14]	[-.02-.22]	14
Commonly used	21	12742	.14	.10	.09	[.09-.18]	[.02-.25]	17
Uncommonly used	8	7696	.04	.08	.07	[-.01-.10]	[-.05-.14]	18
General Ability	31	34289	.18	.07	.06	[.15-.20]	[.10-.25]	20
Commonly used	12	14835	.21	.08	.07	[.17-.25]	[.12-.30]	13
Uncommonly used	19	19454	.15	.04	.03	[.13-.17]	[.12-.19]	55

Note. *k*=number of independent studies; *N*=total sample size; ρ= mean true-score correlation corrected only for sampling error; SD_r=sample-size-weighted observed standard deviation of correlations; SD_ρ=standard deviation of true-score correlations corrected for sampling error; CI=confidence interval around the mean true-score correlation; CV=80% credibility interval; VE= variation in the observed correlations attributable to sampling error.

4.4.3 Moderating Effect of Year of publication

Because the studies included in the database cover a wide time frame (1987–2017), the year of publication was analyzed as a possible moderator variable, three groups were formed for the 30-year period: (1987-1999), (2000-2009), and (2010-2017). This moderator was tested on all TBs categories with no exception. Hence, a total of 18 validities was analyzed for this moderator (six TBs multiplied by three subgroups). As reported in Table 7, across the 18 mean validities, only one validity included zero (a subgroup of 2010-2017 in Controlled Attention). This suggested that TBs with different ability saturations are robust predictors for pilot performance over the years. Examining the 95% CI revealed that only Work Sample TBs did not show overlapping between the three subgroups while the remainder TBs had overlapped subgroups of publications, which gave an indication for insignificant mean differences.

Table 7
Moderating effect of the year of publications (three subgroups)

Type of Test Battery	*k*	*N*	ρ	SD_r	SD_ρ	95% CI	80% CV	%VE
Acquired Knowledge	68	93209	.19	.09	.09	[.17-.21]	[.07-.30]	8
1987-1999	32	49088	.22	.10	.10	[.19-.26]	[.10-.34]	6
2000-2009	10	14360	.19	.08	.07	[.14-.24]	[.10-.28]	11
2010-2017	26	29761	.13	.05	.04	[.11-.15]	[.08-.17]	37
Perceptual Processing	47	48697	.14	.05	.04	[.12-.15]	[.08-.19]	33
1987-1999	13	14225	.12	.04	.02	[.10-.14]	[.09-.15]	67
2000-2009	11	10656	.15	.07	.06	[.11-.19]	[.07-.23]	21
2010-2017	23	23816	.14	.05	.04	[.11-.16]	[.09-.19]	35
Motor Abilities	39	24388	.17	.11	.10	[.13-.20]	[.03-.30]	12
1987-1999	18	11509	.17	.08	.07	[.13-.20]	[.07-.26]	22
2000-2009	16	10600	.13	.11	.10	[.08-.19]	[.00-.26]	12
2010-2017	5	2279	.32	.13	.11	[.21-.43]	[.17-.47]	11
Controlled Attention	29	20438	.10	.10	.09	[.06-.14]	**[-.02-.22]**	14
1987-1999	12	3849	.22	.12	.10	[.15-.28]	[.09-.35]	21
2000-2009	11	9041	.10	.06	.05	[.07-.14]	[.04-.17]	30
2010-2017	6	7548	.04	.07	.06	[-.02-.09]	[-.04-.11]	18
General Ability	31	34289	.18	.07	.06	[.15-.20]	[.10-.25]	20
1987-1999	8	6106	.19	.09	.08	[.13-.25]	[.08-.29]	16
2000-2009	5	10321	.19	.06	.06	[.14-.25]	[.12-.26]	13
2010-2017	18	17862	.16	.05	.04	[.14-.19]	[.10-.22]	33
Work Sample	9	1655	.34	.21	.19	[.21-.48]	[.09-.59]	10
1987-1999	3	554	.55	.17	.16	[.36-.75]	[.35-.76]	9
2000-2009	2	557	.32	.02	0	[.30-.34]	[.32-.32]	100
2010-2017	4	544	.15	.13	.10	[.02-.28]	[.024-.27]	42

Note. *k*=number of independent studies; *N*=total sample size; ρ= mean true-score correlation corrected only for sampling error; SD_r=sample-size-weighted observed standard deviation of correlations; SD_ρ=standard deviation of true-score correlations corrected for sampling error; CI=confidence interval around the mean true-score correlation; CV=80% credibility interval; VE= variation in the observed correlations attributable to sampling error.

When looking closely at validity estimates it appears that the strength of the relationships between the TBs and pilot performance tended to decrease gradually in TBs of Acquired Knowledge, Controlled Attention, and Work Sample. No clear trend was noticed for TBs of Perceptual Processing, Motor Abilities, and General Ability. To gain insight about the types of changes that have occurred over the years, correlation analysis was performed with the year of publication. The results showed negative correlations for TBs of Acquired Knowledge (-.23; *P*>.05), Perceptual Processing (-.06; *P*>.05), General Ability (-.30; *P*>.05), and Work Sample (-.65; *P*>.05), indicating decreases in validity over the years. For Motor Abilities and Controlled

Attention TBs, the correlations were positive (.23 and .13; *P*>.05; respectively) indicating growth over the years. However, none of the tested correlations were significant. Hence, there is little support for Hypothesis 5.

4.4.4 Moderating Effect of Flying Organization

This analysis examined whether the validity of TBs differed depending on the flying organization. Four subgroups were identified: USAF, US Navy, another military, and civilian. There was not enough data to tests this moderator in three instances: subgroups of US Navy in General Ability TBs, US Navy in Work Sample TBs, USAF in Work Sample TBs. Hence, a total of 21 correlations were analyzed for this moderator (six TBs multiplied by four subgroups –3). Table 8 includes the complete results. It was interesting to see all TBs showing significant validities across the 21 relations with no single exception. Concerning mean differences between flying organizations, results of the 95% CI indicated no significant differences between the four subgroups on any TBs. This suggests that all ability saturations of TBs are a robust predictors of pilot performance across settings. The largest mean validity estimates were found for subgroups of the civilian and another military in Work Sample TBs (*r*=.34 for both) while the smallest estimates were found for USAF subgroup in Controlled Attention TBs (*r*=.06). USAF subgroups had the smallest validity estimates among the four subgroups in three occasions: Acquired Knowledge TBs, Motor Ability TBs, and Controlled Attention TBs. Overall, all categories of TBs showed significant criterion-related validity for pilot performance across the four examined subgroups of flight organizations. The pattern of validities among subgroups did not support significant differences between them.

Table 8
Moderating effect of flying organization (four subgroups)

Type of Test Battery	*k*	*N*	ρ	SD_r	SD_ρ	95% CI	80% CV	%VE
Acquired Knowledge	68	93209	.19	.09	.09	[.17-.21]	[.07-.30]	8
USAF	39	65171	.15	.04	.03	[.14-.16]	[.11-.19]	32
US Navy	16	24085	.29	.11	.11	[.24-.35]	[.15-.43]	4
Another Military	8	3335	.18	.08	.06	[.13-.24]	[.10-.26]	37
Civilian	5	618	.15	.06	0	[.10-.20]	[.15-.15]	100
Perceptual Processing	47	48697	.14	.05	.04	[.12-.15]	[.08-.19]	33
USAF	28	44824	.13	.04	.03	[.12-.15]	[.09-.17]	38
US Navy	3	1211	.13	.02	0	[.11-.15]	[.13-.13]	100
Another Military	3	1241	.10	.03	0	[.06-.14]	[.10-.10]	100
Civilian	13	1421	.26	.17	.15	[.16-.35]	[.07-.45]	27
Motor Abilities	39	24388	.17	.11	.10	[.13-.20]	[.03-.30]	12
USAF	10	16478	.12	.07	.06	[.08-.16]	[.04-.20]	14
US Navy	5	2018	.27	.06	.04	[.21-.33]	[.21-.33]	52
Another Military	14	4446	.29	.14	.13	[.22-.36]	[.13-.45]	14
Civilian	10	1446	.17	.11	.08	[.10-.24]	[.07-.27]	53
Controlled Attention	29	20438	.10	.10	.09	[.06-.14]	[-.02-.22]	14
USAF	7	16243	.06	.05	.04	[.03-.09]	[.01-.11]	21
US Navy	7	2572	.23	.10	.08	[.16-.30]	[.13-.34]	26
Another Military	5	501	.29	.10	.05	[.20-.38]	[.23-.35]	80
Civilian	10	1122	.28	.13	.10	[.20-.36]	[.15-.41]	43
General Ability	31	34289	.18	.07	.06	[.15-.20]	[.10-.25]	20
USAF	26	32815	.17	.06	.06	[.15-.20]	[.10-.25]	19
US Navy	0	-	-	-	-	-	-	-
Another Military	2	622	.27	.11	.09	[.12-.41]	[.15-.38]	25
Civilian	3	852	.15	.06	.03	[.08-.22]	[.12-.18]	83
Work Sample	9	1655	.34	.21	.19	[.21-.48]	[.09-.59]	10
USAF	0	-	-	-	-	-	-	-
US Navy	0	-	-	-	-	-	-	-
Another Military	7	1505	.34	.22	.21	[.18-.50]	[.08-.60]	78
Civilian	2	150	.34	.07	0	[.25-.44]	[.34-.34]	100

Note. *k*=number of independent studies; *N*=total sample size; ρ= mean true-score correlation corrected only for sampling error; SD_r=sample-size-weighted observed standard deviation of correlations; SD_ρ=standard deviation of true-score correlations corrected for sampling error; CI=confidence interval around the mean true-score correlation; CV=80% credibility interval; VE= variation in the observed correlations attributable to sampling error.

4.4.5 Moderating Effects of Criterion's level of measurement

This analysis focused on the criterion level of measurement and tested its possible effect on the relations between predictor groups and criterion. Four subgroups were constructed: continuous, ordinal, dichotomous, and contingency. Data were not adequate to perform five analyses from 24 possible analyses (six TBs multiplied by four subgroups). Thus, 19 total analyses was conducted. The results are presented in Table 9. Across different criterion level of measurement, validity estimates of TBs categories were significantly positive in 18 analyses with magnitude ranging from .03 (ordinal/Controlled Attention) to .62 (dichotomous/Work Sample). The criterion subgroup of the ordinal mode in Controlled Attention TBs did not show valid prediction nor did it generalize validity. This was suggested by the inspection of the 95% CI and the 80% CV where both included zero. Inspecting the 95% CI of the four subgroups within each category of TBs indicated that there were not any significant mean differences.

As may be expected, TBs validities based on continuous-scaled criteria demonstrated better prediction of pilot performance than dichotomous-scaled criteria in five categories of TBs. The only exception was the case of Work Sample TBs where it had validity for dichotomous-scaled criteria (only two studies, N=150) larger than continuous-scaled criteria (k = 7, N = 1,505). Similarly, prediction of continuous-scaled performance was better than that of ordinal-scaled performance across four categories of TBs. The exception to this conclusion was observed with Perceptual Processing TBs. The comparison between TBs of ordinal- and dichotomous-scaled criterion resulted in mixed finding. Three instances were in favor of TBs of dichotomous-scaled criterion (Acquired Knowledge, General Ability, and Controlled Attention) and two instances were in favor of ordinal-scaled criterion (Perceptual Processing and Motor Abilities). TBs used contingency table criteria for pilot performance were very few to allow valid assessments (only

two cases both of which had only two studies each). In general, results suggest robust mean validity of TBs categories across criterion levels of measurement.

Table 9
Moderating effect of criterion level of measurement (four subgroups)

Type of Test Battery	*k*	*N*	ρ	SD_r	SD_ρ	95% CI	80% CV	%VE
Acquired Knowledge	68	93209	.19	.09	.09	[.17-.21]	[.07-.30]	8
Continuous	40	37255	.26	.11	.10	[.23-.29]	[.13-.39]	89
Ordinal	6	23188	.13	.02	.01	[.11-.14]	[.11-.14]	59
Dichotomous	18	29878	.15	.03	.02	[.13-.16]	[.12-.18]	49
Contingency	2	1254	.08	.02	0	[.05-.12]	[.08-.08]	100
Perceptual Processing	47	48697	.14	.05	.04	[.12-.15]	[.08-.19]	33
Continuous	24	6235	.13	.08	.05	[.10-.16]	[.07-.19]	64
Ordinal	5	21985	.14	.04	.04	[.11-.18]	[.09-.19]	13
Dichotomous	9	18963	.13	.02	0	[.12-.14]	[.13-.13]	100
Contingency	0	-	-	-	-	-	-	-
Motor Abilities	39	24388	.17	.11	.10	[.13-.20]	[.03-.30]	12
Continuous	14	5604	.23	.09	.08	[.18-.27]	[.13-.32]	28
Ordinal	2	915	.20	.07	.06	[.10-.30]	[.13-.27]	38
Dichotomous	18	16135	.13	.09	.08	[.09-.17]	[.02-.23]	14
Contingency	2	576	.13	.09	.06	[.01-.25]	[.05-.21]	46
Controlled Attention	29	20438	.10	.10	.09	[.06-.14]	[-.02-.22]	14
Continuous	14	4295	.21	.11	.09	[.16-.27]	[.10-.33]	26
Ordinal	2	7240	.03	.05	.05	[-.04-.10]	[-.03-.09]	10
Dichotomous	9	8671	.10	.06	.05	[.06-.13]	[.04-.16]	33
Contingency	0	-	-	-	-	-	-	-
General Ability	31	34289	.18	.07	.06	[.15-.20]	[.10-.25]	20
Continuous	19	6291	.19	.11	.09	[.14-.23]	[.07-.30]	25
Ordinal	3	15707	.16	.02	.01	[.14-.18]	[.14-.18]	50
Dichotomous	8	11715	.19	.07	.06	[.14-.23]	[.11-.26]	14
Contingency	1	-	-	-	-	-	-	-
Work Sample	9	1655	.34	.21	.19	[.21-.48]	[.09-.59]	10
Continuous	7	1344	.28	.14	.12	[.17-.38]	[.12-.43]	23
Ordinal	0	-	-	-	-	-	-	-
Dichotomous	2	311	.62	.21	.20	[.33-.92]	[.36-.88]	5
Contingency	0	-	-	-	-	-	-	-

Note. *k*=number of independent studies; *N*=total sample size; ρ= mean true-score correlation corrected only for sampling error; SD_r=sample-size-weighted observed standard deviation of correlations; SD_ρ=standard deviation of true-score correlations corrected for sampling error; CI=confidence interval around the mean true-score correlation; CV=80% credibility interval; VE= variation in the observed correlations attributable to sampling error.

4.4.6 Supplementary Analysis

As a supplementary analysis, the Hedges and Olkin (1985) approach to meta-analysis was conducted separately for each ability category of TBs using the overall criterion. As shown in Table 9, the estimates of weighted-mean correlations were .39, .25, .23, .18, .17, and .16 for TBs saturated with Work Sample, Motor Abilities, Controlled Attention, General Ability, Acquired Knowledge, and Perceptual Processing, respectively. The 95% CIs indicates that we can be quite confident the true average validities of all TBs are positive, exceeding zero. The *Q*-within tests were significant, and the I^2 statistics showed that most of the observed variability in validity estimates were due to true differences, not sampling errors. Thus, further moderator analyses are justified. The overall conclusion of this approach appeared consistent with the former Schmidt and Hunter (2015) approach. However, two important differences need to be highlighted here. First, higher magnitudes of validities were noted for TBs of Motor Abilities and Controlled Attention. The validity estimates increased from.17 to .25 for Motor Abilities TBs and from .10 to .23 for Controlled Attention TBs. More importantly, in addition to the higher magnitude noted for Controlled Attention TBs, the 95% CI around its mean was far from zero [.18–.28], which differs largely from the previous result that found a small validity estimates of .10 with barely non-zero 95% CI [.06–.14] and zero 80% CV [-.02–.22]. This supports the validity of this predictor as pilot performance.

Table 10
Reanalyzing TBs-criterion relationships using Hedges and Olkin's approach and testing publication bias

	k	*N*	$\bar{r}$	SD_r	95% CI	z	I^2	*Q*	Rank *r*	R-Z	Fail-safe *N*
AK	68	93209	.17	.01	[.15-.20]	14.6**	89.8	901.2**	.16	-.46	44144**
PP	47	48697	.16	.02	[.13-.19]	9.6**	87.6	160.8**	.37**	4.53**	9317**
MA	39	24388	.25	.02	[.20-.29]	10.7**	89.9	359.0**	.16	2.35*	1083**
CA	31	34289	.23	.03	[.18-.28]	8.6**	88.6	226.6**	.04	2.33*	3506**
GA	29	20438	.18	.02	[.15-.22]	9.9**	87.1	157.4**	-.11	-.41	7524**
WS	9	1655	.39	.09	[.21-.57]	4.2**	92.1	125.2**	.11	-.33	725**

Note. *k*=number of independent studies; *N*=total sample size; $\bar{r}$= average validity from the random effects meta-analysis; SD*r*= standard deviation of the validity corrected for sampling error; CI=confidence interval around the mean correlation; I^2=percentage of variance beyond sampling error; Q=chi-square test for homogeneity of observed validities; Rank r= Rank correlation test for funnel plot asymmetry (Kendell's Tau); R-Z=Regression test for funnel plot asymmetry; Fail-safe N= the sample size would have to exist to bring the difference in validity between predictor and criterion down to ρ= .05; AK=Acquired Knowledge; PP=Perceptual Processing; MA=Motor Abilities; CA=Controlled Attention; GA=General Ability; WS=Work Sample.
** *P*<.01, * *P*<.05

4.4.7 Publication Bias

The evaluation of publication bias revealed different conclusions from different analyses. As shown in Figure 5, most of the funnels plots are not visually symmetric. Deviations from funnel-shaped distribution can be observed in several TBs' categories such as Perceptual Processing and Motor Abilities. The symmetric shape may be somewhat detected for Acquired Knowledge and General Ability TBs. Statistical tests of funnel plots using Begg rank correlation indicate that all TBs categories were nearly symmetric, with Perceptual Processing being the only exception. Thus, there was no evidence of publication bias for five of six TBs. Based on Egger weighted linear regression test, three types of TBs found to be symmetric (Acquired Knowledge, General Ability, Work Sample) and three TBs found to be asymmetric (Perceptual Processing, Motor Abilities, Controlled Attention). Thus, there were evidence of publication bias for three TBs but no evidence for the other three. Last, the fail-safe *N* analysis for each TBs' predictor indicated that many null studies would need to be located for the two-tailed *p*-value to exceed 0.05. Specifically, fail-safe *N* was 44,144 for Acquired Knowledge TBs, 9,317 for Perceptual Processing

TBs, 7,524 for Motor Ability TBs, 3,506 for Controlled Attention TBs, 1,083 for General Ability TBs, and 725 for Work Sample TBs. This means that there would have to be from 725 (Work Sample) to 44,144 (Acquired Knowledge) null studies included in the present study with zero difference in validities to bring the difference in validities down to insignificant magnitudes. According to the significant results established by this test, effect sizes are not expected to be confounded by publication. For the sake of completion, forest plots parallel to funnel plots are shown in Figure 6.

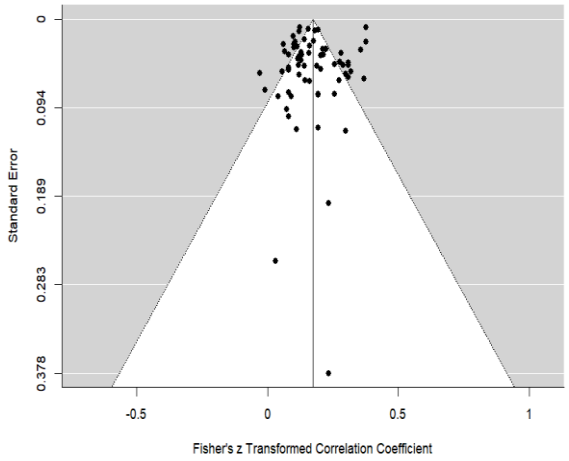

Figure 5 (a). Funnel plot of Acquired Knowledge TBs

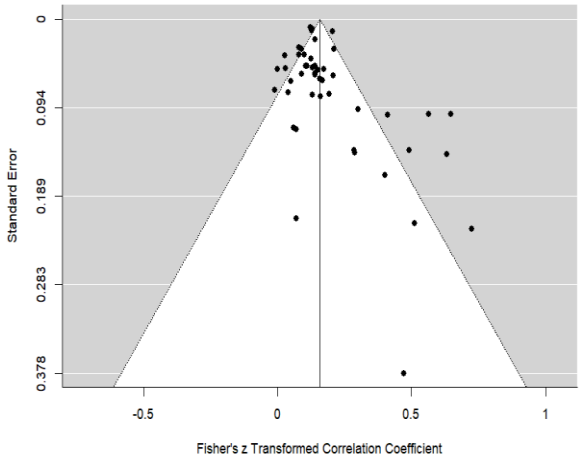

Figure 5 (b). Funnel plot of Perceptual Processing TBs

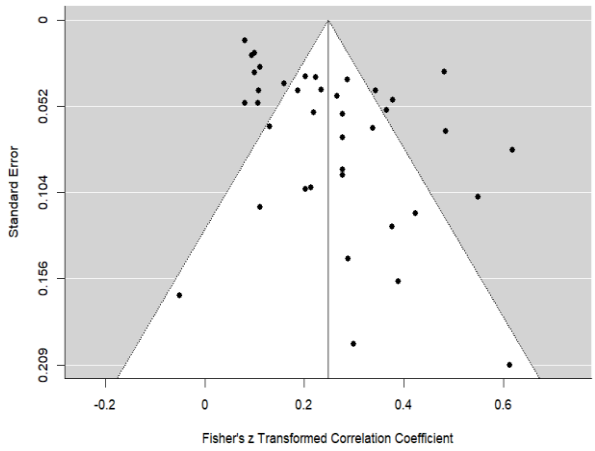

Figure 5 (c). Funnel plot of Motor Abilities TBs

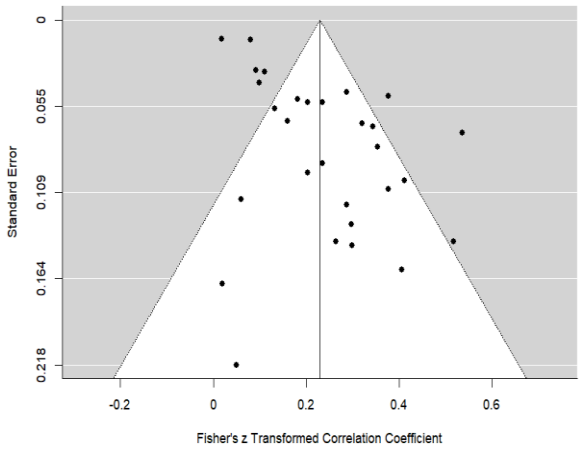

Figure 5 (d). Funnel plot of Controlled Attention TBs

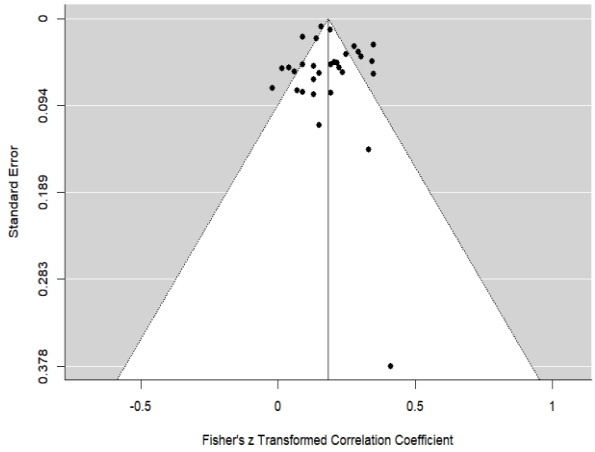

Figure 5 (e). Funnel plot of General Ability TBs

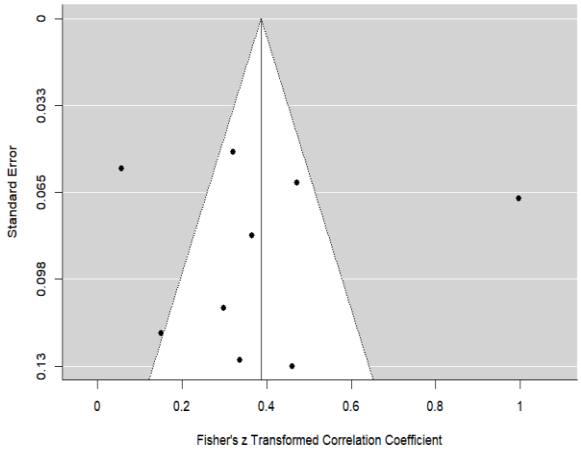

Figure 5 (f). Funnel plot of Work Sample TBs

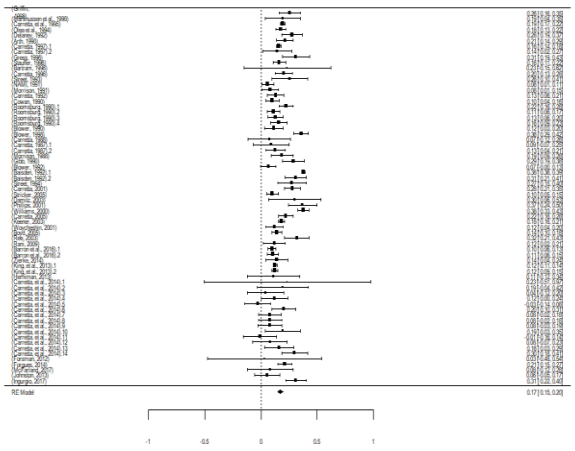

Figure 6 (a). Forest plots of Acquired Knowledge TBs

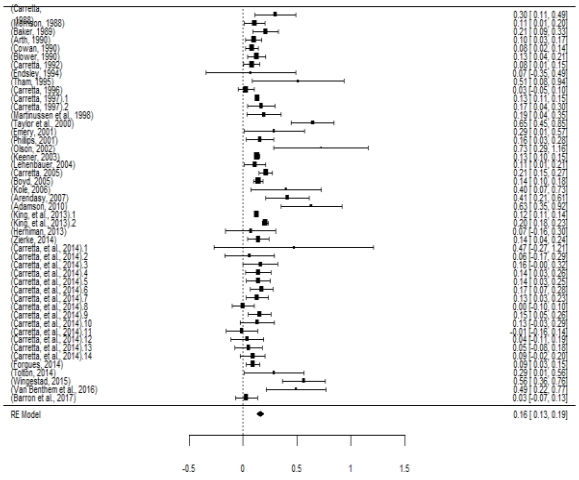

Figure 6 (b). Forest plots of Perceptual Processing TBs

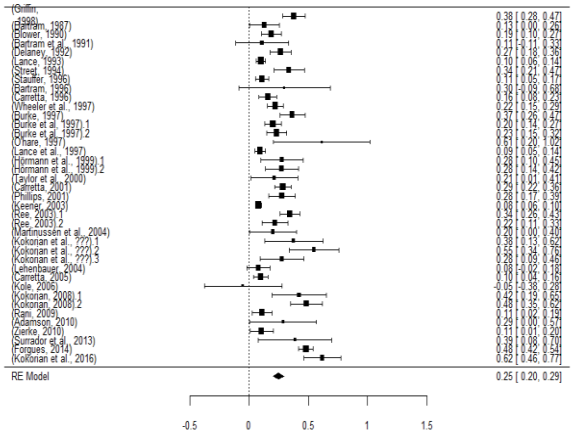

Figure 6 (c). Forest plots of Motor Abilities TBs

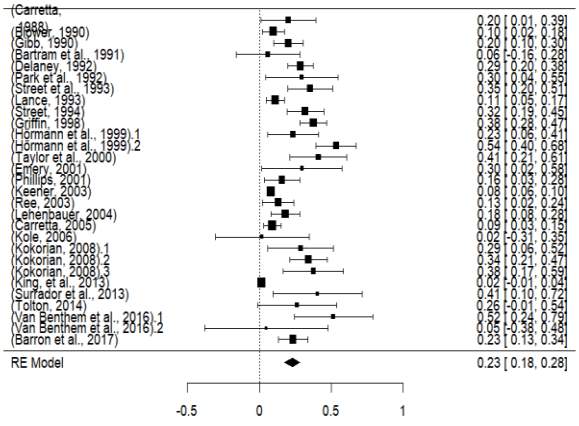

Figure 6 (d). Forest plots of Controlled Attention TBs

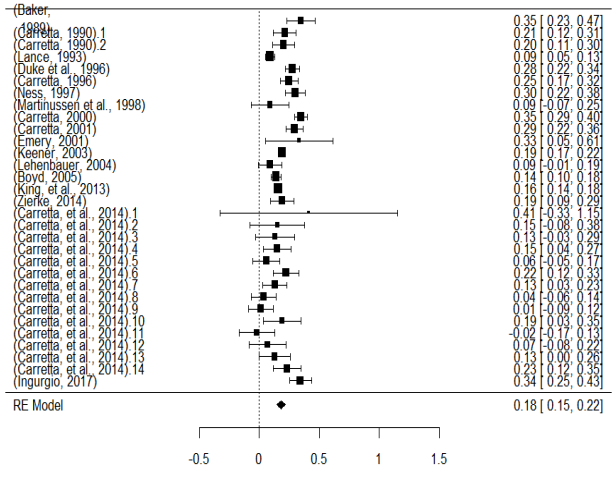

Figure 6 (e). Forest plots of General Ability TBs

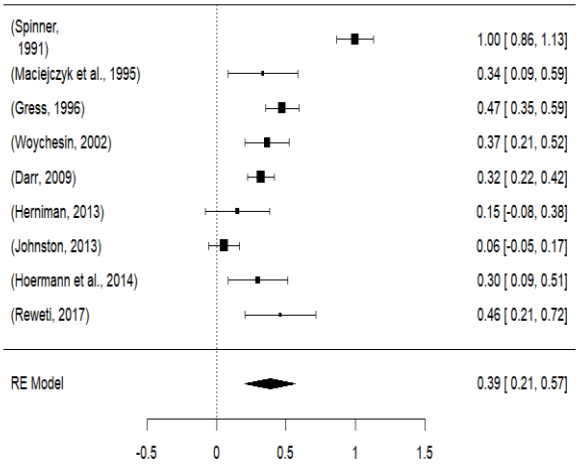

Figure 6 (f). Forest plots of Work Sample TBs

CHAPTER 5. DISCUSSION

This meta-analysis has provided the most comprehensive statistical review of the relationship between composite scores of test batteries and pilot performance criteria to date. The aim was to determine the mean validity of six distinct types of ability batteries for predicting pilot performance across four specific criteria (flying performance rating, graduate/attrite training, academic performance grade, flight simulator performance rating), as well as one global criterion. Additionally, five variables were examined as possible moderators influencing predictor–criterion associations (number of test in the battery, regularity of TBs use in pilot selection, the year of publication, the type of flying organization, the scale/type of criterion). This examination represents an extension of earlier meta-analyses conducted more than two decades ago assessing the criterion-related validity of single-construct psychological tests as predictors for pilot performance. The categorization schema of TBs utilized in this study based on broad ability constructs is novel, and this is one of the only occasions in which ability battery composite scores have been subjected to verification in a meta-analysis context. It was clear that different ability saturations of TBs identified primarily on the basis of the CHC model can provide practically useful levels of statistical prediction of several important criteria of flight performance. Despite the small magnitude of mean validities, the six categories of TBs correlated meaningfully with the criteria of pilot performance, and five of them showed a possible generalization of validity across samples and organizations. The idea that some abilities would play a greater role in pilot performance than others has been confirmed by the fact that different categories of TB ability saturations yielded different mean validities. Researchers and practitioners, therefore need to take into consideration the broad ability saturation underlying each computed TB composite score.

Findings of this study thus inform the debate surrounding the effectiveness of ability batteries for predicting pilot performance.

5.1 Prediction of Four Pilot Performance Outcomes

The investigation of the criterion-related validity of six broad ability constructs of TBs for four outcomes of pilot performance supports the validity of TBs and establishes their importance as predictors of pilot performance. Out of 21 examined relations between TBs and criteria, only two relations related to Controlled Attention TBs had some validity issues (*with* flying rating and academic grade). The remaining 19 relations gave significant evidence for predicting pilot performance. Differences are noted in the magnitude and the respective 95% CI and 80% CV. Overall, criteria of academic grade and simulator rating were associated with higher mean correlations as compared to flying rating and graduate/attrite training. The academic grade was the best-predicted criterion by TBs of Acquired Knowledge, Perceptual Processing, and General Ability. Findings of previous meta-analyses suggested that initial ability scores are more valid predictors of training success criteria than criteria based on subsequent job performance (Hirsh et al. 1986; Hunter, 1986; Hunter & Hunter, 1984; Pearlman et al., 1980). Academic performance in flight training is probably the closest criteria for training success in other professions, and hence the present finding is in consistent with previous studies finding. Also, simulator rating was the best-predicted criterion by TBs of Motor Abilities and Controlled Attention. The relatively high validity noted for these two ability batteries may be understood considering the nature of their tests/tasks that require competencies like those needed for piloting ability (e.g., speed of processing, attention, time-sharing, compensatory tracking; e.g., Gibb & Dolgin, 1989; Johnston, 1996; Taylor et al., 1994; Tsang & Shaner, 1998).

Another notable finding is the lower validity magnitude of TBs' associations with graduate/attrite training criterion as compared to flying performance rating criterion (*M*= .15 vs. *M*= .19). This trend is somewhat expected given the typical dichotomized scale of graduate/attrite criterion which tends to "underestimate" correlations coefficients (Cohen, 1983). Some interesting findings were observed concerning the variability in validity magnitudes for a the different types of TBs. For example, for the flying rating criterion, TBs of Motor Abilities and General Ability demonstrated the second highest predictive validities ($\bar{r}$=.22 and .17, respectively). Same result was also noticed for both TB types using the graduate/attrite training criterion, with mean validity of .14 for both. Additionally, due to the nature of the constructs of Acquired Knowledge and Perceptual Processing, it was projected that Acquired Knowledge TBs will predict academic grades better than other criteria, and Perceptual Processing TBs will predict flying rating better than other criteria. Results showed that both types of ability batteries had as the cademic grade as their best-predicted outcomes. This finding may be understood when considering the high educational prerequisites required for selecting pilots (e.g., GPA, standardized test scores), which increase their likelihood to be successful academically, regardless of their ability orientations.

The findings also suggest that the academic performance criterion might not be sensitive enough for detecting the different types and levels of cognitive abilities that are essential for pilots, and there is need to include different specific criteria of pilot performance to understand the TBs' relative utility. Controlled attention TBs were found to be a more effective predictor for pilot performance than criteria of graduate/attrite training and simulator rating and, to some extent, academic grade. Unstable findings for this category, however, could be linked to the possibility that TBs included in this category were more diverse and wider in scope, which may increase variability around the mean. Overall, results were consistent with studies emphasizing the role of

criterion type in validation studies, and those stressing the importance of balancing predictor–criterion relations such that broader predictors are matched with broader criteria and specific predictors are matched with specific criteria (Ones & Viswesvaran, 1996; Reeve et al., 2015).

5.2 Prediction of Overall Pilot Performance Outcome

The mean validities of TB ability categories predicting a single global index for pilot performance ranged from .10 (Controlled Attention TBs) to .34 (Work Sample TBs), all of which had 80% CV greater than zero except for Controlled Attention. Three groups of predictors may be identified according to the magnitude of validity estimates: the highest validity group containing TBs of Work Sample (.34), the medium validity group containing TBs of Acquired Knowledge, General Ability, and Motor Abilities (.19, .18, and .17, respectively), and the lowest validity group containing TBs of Perceptual Processing and Controlled Attention (.14 and .10, respectively). Similarities were noted between the three types of TBs in the medium validity group. Because flight simulators are often associated with high costs and advanced technology, selection tests that are dominated by one of these three broad abilities are more likely to be acquired and attained. Psychomotor ability tasks also tend to rely on advanced equipment such as apparatus and computer-based instruments, which may be not attainable for some organizations. Hence, the TBs of Acquired Knowledge and General Ability become the more accessible type of batteries. Fortunately, these two types of ability are the most frequently used and employed in pilot selection batteries. The current finding gives further support for the continuing utility of these types of TBs, which have at least modest predictive power.

Concerning the smaller magnitude of correlations for Perceptual Processing TBs, it may be linked to the high cognitive demand required by ability batteries of these types as well as the multidimensional scope of measurement covered by these TBs with narrow abilities and process,

which may impact participants' performance. Many neurocognitive tests batteries consist of a large number of tests/tasks that measure wide-ranging constructs of cognitive functioning (e.g., King et al., 2011; King et al., 2013). Nevertheless, TBs of Perceptual Processing is still not that distant from the "medium validity" group, with correlation estimate supporting their validity generalization. The finding regarding TBs of Controlled Attention ($\bar{r}$=.10; CV= [-.02-.22]) is especially noteworthy because many primary studies have shown that this category (i.e., multitasking tasks) is important predictors of pilot performance (e.g., Barron, & Rose, 2017; Hoover, & Russ-Eft, 2005; Morgan et al., 2013). The inconsistent result noted for this group in the current study may be attributed to the mixed measures of dual tasks that were included in this category, with no attempt to further sorting them out to a more specific type of multitasking (e.g., cognitive dual tasks, psychomotor dual tasks, ca ombination of perceptual-motor tasks). Another possibility is that an outlier value exists in the data that may have influenced the validity estimate of Controlled Attention TBs. The inspection of the forest plot of this category suggests that the King et al. (2013) primary study may be the most influential outlier. It is recommended that more single studies on the validity of Controlled Attention TBs to be conducted for specific combinations of abilities so as to allow further categorization of constructs and processes included in these batteries.

Overall, findings based on a single global criterion lead to positive conclusions about the validity of the different types of TBs in predicting pilot performance, with possible generalizability. Although progress has been made in understanding the validity of cognitive abilities for predicting pilot performance, results of meta-analyses have shown that the observed validities of cognitive abilities are modest (in the .10s to .30s range) at best (Hunter & Burke,

1994; Martinussen, 1996). Results of this study appear consistent with this overall finding, with different conclusions realized for the validity of each type of TB.

5.3 Flight Simulator as both a Predictor and a Criterion

As noted earlier, the flight simulator score (i.e., work sample) was found to be the best predictor across three specific criteria (flying rating, graduate/attrite training, and academic grade) and one overall criterion of pilot performance. With mean validities of .35 for flying performance, .34 for graduate/attrite training, .24 for academic performance, and .34 for the overall index, and the respective 80% credibility values lying clearly above zero, the conclusion that validity generalization of this predictor exists across different samples and situations is supported. The robustness of flight simulator as a predictor can be realized across four criterion-based meta-analyses, and seven moderator-based meta-analyses, where it yielded the largest validity estimates among other predictors, with 80% credibility values greater than zero. The findings suggest that simulator is the best "stand-alone" predictor of pilot performance among TB categories. This result is consistent with Carretta and Ree's (2003) conclusion, which identified flight simulators as a strong predictor of pilot performance. Also, it matches the suggestion of Schmidt and Hunter's (1998) review, which established that intelligence measures and work sample tests, as well as their combination, are the best predictors for job performance.

Considering flight simulator as a criterion, the analysis showed that that validity estimates of different TBs categories for flight simulator rating were generally of high magnitude. This is especially true for TBs of Perceptual Processing, Motor Abilities, and Controlled Attention, but it is true to a lesser extent for TBs of Acquired Knowledge. However, it should be noted that all studies that used flight simulator as a criterion were concurrent validity studies, and involved only

experienced pilot samples, not student trainees. A possible explanation for the notable high validity in simulator-based studies is the testing condition of simulation, which is not expected to be as of the actual flying situation in terms of difficulty, complexity, and stress involved. Simulation tasks tend to be easier, less complex, and involve less stress than the actual flying tasks. Hence, such result supports the increased use of simulation in validation studies and recognize simulation as a criterion that has the potential to provide a sound index of flying performance, especially as simulation technology advances. Nevertheless, the financial and training factors that discourage using flight simulator performance as a criterion are reasonably understood.

It is important to note, however, that the present meta-analysis of flight simulator as a predictor was based on only three to six independent samples across the specific criterion and total of nine samples for the global criteria (*N*= 1,655). Similarly, the meta-analysis of flight simulator as a criterion was based on only two to nine independent samples for each type of TB with sample sizes ranging from 232 to 1,634. The interpretation of results, therefore, needs to be taken with caution.

5.4 Number of Tests in the Battery

Analyzing the validity of TBs across different subgroups provided further support for their validities as significant predictors for pilot performance. Based on the number of tests in the battery, the two constructed subgroups (small battery/large battery) produced small to moderate mean validities across three types of TBs and demonstrated generalized validity. Comparing the magnitude of mean validities obtained from the heterogeneous tests in this study with those obtained from homogenous tests in the previous meta-analyses reveals some differences. For example, TBs of Motor Abilities in the present study showed mean validity of .26 and .12 for large battery and small battery subgroup respectively, both of which are values below the validity of .32

reported in Hunter and Burk (1994) for Gross dexterity (an index of psychomotor ability) and around the mean validity of .20 reported by Martinussen (1996) for a combination of psychomotor and information processing. The heterogeneous tests examined here for composite scores of multiple tests may give more accurate estimates for the higher order Motor Abilities saturating TBs whereas homogenous tests examined in previous analysis from ability-specific tests may give more accurate estimates for that specific abilities. Practitioners may trust that TBs with higher numbers of tests are more effective at predicting performance than those with lower numbers of tests due to the misbelief that more is better. The results of the current meta-analyses on large samples counter this belief. It was contrary to the expectation, to find in the present study that small battery subgroup showed higher mean validity than the large battery group in two types of TBs. This result suggests being more mindful about not only the number but also the content representativeness of tests when designing ability batteries. Clearly, performing many additional tests in the battery does not necessarily improve predictions about pilot performance.

5.5 Regularity of TB Use in Pilot Selection

The regularity of a TB's use in pilot selection was analyzed as a possible moderating variable. The results indicated significant mean validities and significant 80% credibility values across TBs types for both constructed subgroups (commonly used, uncommonly used) except for one case. This finding supports TBs' validities irrespective of the direct purposes underlying their designs and constructions. As hypothesized, the commonly-used TBs in pilot selection yielded higher mean validity than the uncommonly-used TBs. It was interesting to see Perceptual Processing TBs contrasting this pattern of expected validities. Among the five examined categories of TBs, only the subgroup of uncommonly used in Controlled Attention had a mean validity of .05 fall in the 95% CI and the 80% CV that included zero. These results imply that flying organizations

may be encouraged to utilize the available ‘off the shelf’ batteries for pilot selection when specialized batteries are not available. For the unexpected finding related to Perceptual Processing TBs, a possible explanation is a difference in sample size between the two subgroups (*k*= 9 for commonly used, *k*= 38 for uncommonly used). Adding more studies of commonly used TBs of Perceptual Processing may provide a fairer comparison with the quality of uncommonly used TBs. Moreover, the continued efforts being made to improve the Perceptual Processing domain to incorporate the recent advancement in intelligence models (Hoelzle, 2008), especially those linked to CHC theory (Alfonso et al., 2005), may have yielded newly-designed, and so less commonly-used, tests with high predictive power.

5.7 Year of Publication

Because the studies collected for this meta-analysis covered nearly 30 years of research, the expectation was to detect some changes in validity estimates. Based on three created subgroups of the year of publication, the trend was not clear enough to draw a firm conclusion about TBs’ validities over time. Although TBs of Acquired Knowledge, Controlled Attention, and Work Samples did show some decline in mean validities over year, the differences between the three decades were not significant. This trend became clearer when correlation analysis was used to assess relationships. Results showed negative correlations between the year of publication and validity coefficients of TBs of Acquired Knowledge, Perceptual Processing, General Ability, and Work Sample, and positive correlations with TBs of Motor Abilities and Controlled Attention. As discussed by Martinussen (1996) who also noticed a similar negative trend, this finding does not necessarily indicate a true decline in criterion-related validity but might be an effect of pilot selection procedure changes over the years. Such trends can also be attributed to the reduction of the selection ratio or the variation among applicant populations over the years. Because this study

did not attempt to correct for range restriction, it is difficult to explain the noted decline in the validity of these scores' use for pilot selection over the years.

5.6 Flying Organization

Subgroup analysis of flying organization shows that different saturation categories of TBs are valid predictors across organizations with no observed exception. By the 80% CVs, there is empirical evidence to conclude that there is validity generalization of TBs for predicting pilot performance across four distinct services (USAF, US Navy, another military, civilian). The magnitude of the coefficients differed between the four flying organizations. Apparently, USAF subgroup had lower validity estimates than for other subgroups in at least three categories of TBs (Acquired Knowledge, Motor Abilities, Controlled Attention). The subgroup of another military, on the contrary, had higher validity estimates in three categories of TBs (Motor Abilities, General Ability, Controlled Attention). Comparison between the two US services revealed that TBs of General Ability and Acquired Knowledge were the best predictor for USAF pilot performance ($\bar{r}$= .17 and .15, respectively) while Acquired Knowledge and Motor Abilities ($\bar{r}$= .29 and .27, respectively) were the best predictor for US Navy/Marine pilot performance. However, the magnitude of the effect sizes generally suggests these tests are more useful for screening US Navy/Marines than Air Force pilots. Additionally, Perceptual Processing TBs appeared to be an especially important predictor for pilot performance within civilian organizations ($\bar{r}$= .26) as compared to other flying organizations. Nevertheless, as some correlations were obtained from small numbers of studies (e.g., 2, 6), related findings should be interpreted with caution. Overall, there is evidence for validity generalization for the ability batteries across organizational settings. This is impan ortant conclusion supports the main finding, giving it a further degree of constancy.

5.7 Criterion Level of Measurement

Subgroup analysis of this moderator indicated that the continuous-scaled criterion mode was predicted better than the dichotomous- scaled or ordinal-scaled criterion modes by four types of TBs (Acquired Knowledge, Motor Abilities, Controlled Attention, and General Ability). One exception was in favor of an ordinal-scaled criterion (Perceptual Processing TBs), and another exception was in favor of a dichotomous-scaled criterion (Work Sample TBs). By the 80% CVs, it is possible to conclude that across 23 TBs-criterion' mode relations there is evidence for validity generalization of TBs' categories for predicting pilot performance. Only the subgroup of ordinal scale criteria that was predicted by Controlled Attention TBs deviated from this general effect (N= 2, $\bar{r}$= .03). This suggests that TBs are robust predictors of pilot performance across criterion levels of measurement. The results support the conclusion that different criterion scaling yielded different magnitudes of correlations with the TB predictor scores. However, sa tatistically meaningful moderating effect caused by this variable cannot be asserted.

5.8 Supplementary Analysis

Utilizing two meta-analysis approaches in this study supports the overall conclusion and complements the findings. Both are commonly-used approaches of meta-analysis in organizational psychology. The overall results support the criterion-related validity of the six types of TBs. The lack of generalizability noted for TBs of Controlled Attention based on the Schmidt and Hunter (2015) approach was met with a high magnitude of validity coefficient and a 95% CI clearly distanced from zero based on the Hedges and Olkin (1985) approach. The notable changes in validity estimates of Controlled Attention TBs presents an example for the potential strength and weakness in meta-analysis approaches, and the importance of combining more than one approach in the investigation. It is not uncommon to observe conservative estimates of effect size by the use

of Schmidt and Hunter (2015) approach, which was documented by some comparative studies as an impending drawback for this approach (e.g., Johnson et al., 1995). In contrast, the strength of estimating parameters using inverse variance weights as Hedges and Olkin (1985) approach was advantageous in this case. Applying this method provides weights that directly specify the degree of precision due to uncertainty attributable both to sampling error and to underlying variability in the population effect sizes (Brannick et al., 2011). Regarding publication bias, the multiple techniques used for the analyses indicated mixed findings. Taken together, the results revealed a reasonable degree of unbiased publication, supportive of the TBs' validities.

CHAPTER 6: IMPLICATIONS, LIMITATION, AND CONCLUSIONS

6.1 Practical Implications

In view of the limited number of meta-analyses in aviation, the reliance has been mostly on in-house validity studies or findings from other primary studies that may be only partially relevant, assuming generalizability to any context of pilot selection and assessment. The findings presented here are useful because they provide support for the use of at least five types of TBs across different settings, based on a large-scale meta-analysis of primary validity studies. Large samples were used in the current analysis (based on the overall criterion) ranging from 20,438 to 93,209 for the typical tests batteries and 1,655 for Work Sample tests, which are exceptionally high to the extent that allows drawing trustworthy conclusions. The results indicated that six categories of test batteries are valid predictors for pilot performance and five of them generalize validity. This finding supports the evidence for the use of TBs in pilot selection regardless of service being selected for, the organization, or the country. More importantly, the results support the use of TBs for pilot selection with varying degree of validity attributed to the ability saturation dominated the batteries.

In general, the six broad ability saturations showed a fair degree of validity which evidenced from the meta-analysis of subgroups across different moderator variables. For the overall criterion, the rank of TB predictors as per the mean validities was as follows: Work Sample (.34), Acquired Knowledge (.19), General Ability (.18), Motor Abilities (.17), Perceptual Processing (.14), and Controlled Attention (.10). Although five of these estimates are considered small/low effect size ($\bar{r}$= below than .25; Lipsey & Wilson, 2001), there are important inferences may be gained from the results. The implication of this finding is to increase the attention given to

flight simulator as it showed the best single predictor of pilot performance. The second group of predictors that should be maintained and continued to be emphasized when designing selection tests is those saturated with Acquired Knowledge, General Ability, and Motor Abilities. For the third group containing TBs of Perceptual Processing and Controlled Attention, their utilities are also evidenced, and they may be employed as additional predictors for pilot performance.

Another finding with practical implications was determined by moderator analyses. Testing the number of tests in the TBs revealed that TBs containing two to four tests performed better than those containing five or more tests. Because a large number of tests incorporated in the battery did not greatly improve prediction, the focus should be directed to the quality and content of the tests rather than quantity. This should help to reduce the costs typically associated with constructing and maintaining psychometric measures. Also, the result regarding the validity of some TBs that are uncommonly used in pilot selection supports their use in pilot selection and provides an effective solution for organizations that do not have in-house designed batteries or those that want additional measures of pilot abilities. Another finding with practical implications is that, in all cases, the magnitude of the validity was larger in flying performance rating studies than in studies that used graduation or attrition from training as a binary outcome. Consequently, the effect of performance criterion type in moderating the TBs' relations with outcome criteria necessitates being thoughtful about the most appropriate criteria for a given validation study. Moreover, these findings highlight the importance of meta-analysis in aviation psychology and provide positive evidence for the continued and expanded use of TBs for pilot selection. It extended our understanding of the impact of the broad ability saturations of TBs on the prediction of pilot performance across different criteria. The present results also suggest that for specific TBs, additional factors may also serve to moderate the criterion-related validity. Thus, additional

research will be needed to examine further the factors that moderate criteria-related validity of TBs.

6.2 Limitation

Despite the notable strengths of this study, several limitations exist that need to be addressed. First, the study might be biased in that it included only primary studies that were published in the English language. There might be important primary studies published in another language than English that would have contributed to the overall finding if they were viable for inclusion. However, the fact that this study included data from more than 89 primary studies minimize this bias and gives the present findings some degree of practical generalizability. Second, the results were not corrected for measurement artifacts, which certainly left wide room for variance to be explained. Future meta-analyses are encouraged to correct for most relevant artifacts. Related to this point, one reason for disregarding the correction of measurement artifacts in this study was the lack of information provided by the primary studies that would facilitate the correction process (e.g., range restriction ratio, predictor reliability, criterion reliability, or dichotomization ratio). It is suggested that future primary validity studies include complete information about the psychometric properties of the measures and criteria. Additionally, even with the attempt in this study to examine four important criteria for pilot performance, future studies should explore even more specific criteria to see whether the pattern of validity coefficients also holds for more specific outcome criteria (e.g., extra flying hours, mishaps, attrition from service with specific reasons, career progress).

Third, some moderator analyses were based on a relatively small number of studies, which may bias the mean validity estimates. Some subgroups may have been underrepresented with sample sizes as low as two or three, or in some cases not even represented. Additional primary

studies are recommended to be carried out to supply future meta-analysis with more comprehensive data. Fourth, the categorization of TBs was based on the broader abilities proposed by CHC theory. Even with the strong psychometric basis underlying the structure of the abilities in the model, the four "big" factors used in this study were a practical operationalization and are still in need of rigorous empirical investigation. Also, categorization was based on qualitative inspection and the thorough understanding of constructs saturating the test battery. A quantitative-based classification method could provide a stronger basis for categorization of batteries into construct saturation types.

Fifth, another limitation is that the presence of military samples was much higher than that of civilian samples (N= 19), which may raise a concern regarding the degree to which results generalize to civilian settings. However, the moderator analysis result about flying organization, in fact, supports the validity generalization of TBs across organizations. Sixth, a final limitation is that some other potential moderators were not considered in this study (e.g., sample nationality, aircraft type, administration format of the battery, gender representation, mean age). Testing for such moderators may increase the explained variance and strengthen the validity generalization concluded from the study. Counter to these limitations; the present meta-analysis also has several strengths. First, this study is one of the first attempts to verify the generalizability of criteria-related validity of multiple tests batteries for pilot selection. It is also one of the first attempts to link test batteries with CHC theory in a meta-analytic investigation. Second, the present meta-analysis benefited from recent meta-analytic practices in industrial/organizational psychology research as it utilized the concept of construct saturation for investigating the complex and multidimensional constructs underlying composite of test batteries. Also, it applied two widely-used approaches of meta-analysis to inform the decisions made throughout this investigation. Third, the primary

studies collected for this meta-analysis included TBs with several broad ability saturations, used in many types of flight organizations in several countries across an extended, recent timeframe. The results concluded in this study should be far more precise than those of any single primary study.

6.3 More Recommendations and Research Opportunities

The results of this study can be beneficial to aviation psychology practitioners and researchers by providing further insight and understanding of cognitive abilities that contribute to successful training and flying performance. Aviation organizations may utilize some of the findings provided by this study to lessen attrition rates in flight programs through changes to their selection test batteries. The conclusions of the study may also be a base for through future investigations of specific broad and narrower cognitive ability predictors of ab-initio pilots. The following are some additional recommendations and research opportunities from the present study that also cover implications for practice.

(1) The meta-analysis in this study focused on test batteries saturated with six broad categories of cognitive ability in which five of them mapped with a broad model of CHC theory. No attempt was made to extend the model investigation to include the 16 broad cognitive abilities of Stratum II or the narrower abilities of Stratum I. It is recommended that further meta-analysis is extended to more specific cognitive abilities identified by the CHC model to evaluate their association with pilot performance. Additionally, each category group of the broad ability saturation was examined individually as a predictor for flight performance. Assigning each TB to one broad ability construct may not indicate the full potential of a composite score as a predictor, and may not capture the true value of each TB as an indicator of flight performance. Rather than focusing on single composite scores, it is recommended that future research could validate conclusions about how the cognitive

abilities within the TB group are potentially intercorrelated and work together to predict flight performance.

(2) Prior meta-analytic research on the CHC model-based cognitive abilities as predictors of flight performance has been scarce. This study has indicated the importance of cognitive ability constructs saturating the accumulated test batteries for flight performance. In order to advance our understanding of the role of complex constructs in the prediction of pilot performance, more research with different methodology and scope may be needed to validate the use of these broad abilities with flight training. The present study relied on qualitative judgment to categorize the composite scores to the identified groups of ability saturations. Conclusions regarding the validities of TBs would be more credible if a quantitative method (e.g., factor analysis) were available to determine more reliably the higher-order construct of each TB.

(3) For a given predictor construct, practitioners may consult this study to understand the expected validity of each type of TB as a predictor of pilot performance. For example, the results of this study indicated that the criterion-related validity for the TBs saturated with Acquired Knowledge was higher than that saturated with Motor Abilities. Such knowledge is important for the designers of selection tests and may inform their choice of cognitive abilities to incorporate. Relatedly, TBs of Controlled Attention as a predictor for flight performance require further attention. The current analysis indicated that scores in this group were statistically significant as a predictor of flight performance, but evidence for validity generalization was not sufficient. If studies continued to show significant associations between scores of Controlled Attention TBs and flight performance, that should support the use of this type of battery for pilot selection.

(4) Researchers should be thoughtful about the structure of composite scores of cognitive abilities, and the suitability of particular performance criteria. Different TBs may be most strongly

related to different types of performance criteria, e.g., TBs saturated with Acquired Knowledge would be expected to perform better with future academic grade as a criterion, while TBs saturated with work sample would be expected to perform better with flying performance rating as a criterion. Thus, attention should be paid to the possibility that certain composite scores may have lower validity for predicting some criteria of pilot performance, and that other composite scores might be more valid predictors for these criteria.

(5) It is recommended that meta-analysis researchers make use of the six test score ability categories proposed by this study. It can serve as a common framework to communicate research findings and validity evidence of composite scores derived from test batteries.

(6) Sufficient details of construct information should always be reported. Researchers are encouraged to communicate information about the specific contents of TBs composites as well as intercorrelations with the constructs/subtests. Moreover, the limited information on statistical artifacts limited the possibility to provide more precise estimates of true validity in this study. A better practice for reporting data is recommended, especially statistics related to predictor and criterion variables, and necessary for artifact corrections such as means, standard deviations, reliability coefficients, range restriction ratios, and dichotomization ratios.

(7) The present meta-analysis used four specific criteria for pilot performance as well as one overall pilot performance rating index. Future investigations may approach pilot performance differently through longitudinal designs by examining pilot performance over some time. Assessment of students' progression throughout their training program may generate useful knowledge of the relative importance of cognitive abilities for each stage of training. For example, it could be determined that particular ability is a significant predictor for the initial phase of flight

training while another ability is a more useful predictor at later stages after gaining some experience.

(8) Based on the findings of this study, the administrators of aviation organizations are encouraged to incorporate the most predictive type of TBs in their selection test batteries. Composite scores derived from these TBs could be used to assess the potential for student successes in the flight training program and subsequent performance as an operational pilot. The use of TBs that were validated by this study may even be beneficial for increasing retention in flight training programs.

(9) Another recommendation is addressed for the administrators of the University flight program. It is common for University 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 evidence for universities to consider introducing more relevant screening tests as a criterion for flight program admission.

(10) There is a need for more data. More primary studies on the criterion-related validity of TBs for pilot performance would yield more stable estimates of the mean. Some of the moderator results reported in this study were based on relatively small subsets of studies. Similarly, there is a lack of literature in which the differential prediction of test battery scores for various groups (e.g., sex, age, ethnicity) are examined. Additional research could be conducted to investigate the utility of test batteries for predicting different groups of pilots' flight performance. Validation studies in civilian settings, particularly university flight programs, are highly sought after. The current meta-analysis included studies mainly obtained from military settings with fewer from civilian settings.

It was not possible to include “university flight program” in the moderator analysis of flying organizations due to the inadequate number of studies.

(11) Specific recommendations from a selection standpoint for youth programs targeting flying as a potential career are to (a) emphasize the importance of cognitive abilities as selection criteria for flight programs, (b) determine specific cognitive abilities that serve as significant predictors of pilot performance among young people, (c) familiarize participants with common TBs used in pilot selection, and (d) provide opportunities for training directed to the most wanted cognitive abilities for pilots.

6.4 Conclusion

The aim of the present study was to determine the mean validity of TBs with six distinct broad ability saturations (Acquired Knowledge, Perceptual Processing, Motor Abilities, Controlled Attention, General Ability, Work Sample) for predicting pilot performance across four specific criteria (flying performance rating, graduate/attrite training, academic performance grade, and flight simulator performance rating), as well as one overall criterion. Five variables were also considered as possible moderators and were analyzed individually (based on the number of test in the battery, the regularity of TBs use in pilot selection, the year of publication, the type of flying organization, and the scale/type of criteria used). This examination pursued relevant research questions from both a theoretical and an applied point of view. It represents an extension of earlier meta-analyses conducted more than two decades ago assessing the predictive validity of psychological tests for pilot performance. To my knowledge, the TB broad ability saturations categorization utilized in this study was used for the first time in a meta-analysis context, and this is one of the only occasions in which composite scores from multiple test batteries have been subjected to verification by meta-analysis. The overall conclusion of the research showed that TBs

with several different ability saturations can be valid predictors of pilot performance and that five of them generalize validity across samples and organizations for predicting pilot performance. Results also showed varying degrees of validity among different types of TBs, which contribute to the debate about the relative importance of TBs for predicting pilot performance.

REFERENCES

Aamodt, M. G., & Kimbrough, W. W. (1985). Comparison of four methods for weighting multiple predictors. *Educational and Psychological Measurement*, *45*(3), 477-482.

*Adamson, M. M., Samarina, V., Xiangyan, X., Huynh, V., Kennedy, Q., Weiner, M., ... & Taylor, J. L. (2010). The impact of brain size on pilot performance varies with aviation training and years of education. *Journal of the International Neuropsychological Society*, *16*(3), 412-423.

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior*, *32*(8), 1033-1043.

Alfonso, V. C., Flanagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development and interpretation of cognitive and academic abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 185–202). New York, NY: Guilford Press.

*Arendasy, M., Sommer, M., & Hergovich, A. (2007). Statistical judgment formation in personnel selection: A study in military aviation psychology. *Military Psychology*, *19*(2), 119.

*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, Manpower and Personnel Division.

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, *15*(1), 103-133.

*Baker, L. E. (1989). *The Effect of Higher Education Variables on Cadet Performance during 1987 Light Aircraft Training* (No. AU-ARI-88-9). Maxwell AFB, AL: Air University, Airpower Research Inst.

*Baisden, A. G. (1992). *Gender and performance in naval aviation training*. Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*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.

*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.

*Bartram, D. (1987). The Development of an Automated Testing System for Pilot Selection: The MICROPAT Project1. *Applied Psychology*, *36*(3-4), 279-298.

Bartram, D. (1995). Validation of the Micropat battery. *International Journal of Selection and Assessment*, *3*(2), 84-95.

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

*Bartram, D., & Dale, H. C. A. (1991). Validation of the MICROPAT battery of pilot aptitude tests. In *Advances in computer-based human assessment* (pp. 149-169). Springer, Dordrecht.

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). Brook AFB, TX: Armstrong Laboratory, Manpower and Personal Division.

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, *37*(4), 699-725.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088-1101.

*Blower, D. J. (1992). *Performance-Based Testing and Success in Naval Advanced Flight Training*. (No. NAMRL-1378). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Blower, D. J. (1998). *Probability of Success in Primary Flight Training as a Function of ASTB Scores and API Grades: An Example of the Statistical Inferencing Component of the Pilot Prediction System* (No. NAMRL-1404). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Blower, D. J., & Dolgin, D. L. (1991). *An Evaluation of Performance-Based Tests Designed to Improve Naval Aviation Selection* (No. NAMRL-1364). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods*, *10*(4), 689-709.

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, 13, 173-181.

*Boyd, J. E., Patterson, J. C., & Thompson, B. T. (2005). Psychological test profiles of USAF pilots before training vs. type aircraft flown. *Aviation, space, and environmental medicine*, *76*(5), 463-468.

Brannick, M. T., Yang, L. Q., & Cafri, G. (2011). Comparison of weights for meta-analysis of r and d under realistic conditions. *Organizational Research Methods*, *14*(4), 587-607.

*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.

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.

*Carretta, T. R. (1987a). *Field dependence-independence and its relationship to flight training performance* (No. AFHRL-TP-87-36). Brooks AFB, San Antonio, TX: Air Force Human Resources Laboratory.

*Carretta, T. R. (1987b). *Spatial ability as a predictor of flight training performance* (No. AFHRL-TP-86-70). Brooks AFB, San Antonio, TX: Air Force Human Resources Laboratory.

*Carretta, T. R. (1988). *Relationship of Encoding Speed and Memory Tests to Flight Training Performance* (No. AFHRL-TP-87-49). Brooks AFB, San Antonio, TX: Air Force Human Resources Laboratory.

*Carretta, T. R. (1990). Cross validation of experimental USAF pilot training performance models. *Military Psychology*, *2*(4), 257.

Carretta, T. R. (1992). Understanding the relations between selection factors and pilot training performance: Does the criterion make a difference?. *The International Journal of Aviation Psychology*, *2*(2), 95-105.

*Carretta, T. R. (1992). *Predicting Pilot Training Performance: Does the Criterion Make a Difference?* (No. AL-TP-1991-0055). Brooks AFB, San Antonio, TX: Air Force Armstrong Laboratory.

*Carretta, T. R. (1997). Group differences on US Air Force pilot selection tests. *International Journal of Selection and Assessment*, *5*(2), 115-127.

*Carretta, T. R. (2000). *US Air Force pilot selection and training methods* (No. AFRL-HE-WP-TR-2000-0122). Wright-Patterson AFB, OH: Air Force Research Laboratory, Human Effectiveness Directorate.

*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. (2011). Pilot candidate selection method. *Aviation Psychology and Applied Human Factors*. *Aviation Psychology and Applied Human Factors*, 1, 3–8.

Carretta, T. R., King, R. E., Ree, M. J., Teachout, M. S., & Barto, E. (2016). Compilation of Cognitive and Personality Norms for Military Aviators. *Aerospace medicine and human performance*, *87*(9), 764-771.

*Carretta, T. R., & Ree, M. J. (1994). Pilot-candidate selection method: Sources of validity. *The International Journal of Aviation Psychology*, *4*(2), 103-117.

*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).

Carretta, T. R., & Ree, M. J. (2000). *Pilot selection methods.* (No. AFRL-HE-WP-TR-2000-0116). Wright-Patterson, OH: Air Force Research Laboratory, Human Effectiveness Directorate.

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., 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.

Castaneda, M. A. (2007). *A Big Five Profile of the Military Pilot: A Meta-analysis* (Doctoral dissertation). Retrieved from http://etd.fcla.edu/WF/WFE0000097/Castaneda_Michael_Anthony_200712_MA.pdf.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*(3), 153.

Clark, V. L., & Kruse, J. A. (1990). Clinical methods: the history, physical, and laboratory examinations. *Jama*, *264*(21), 2808-2809.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*(1), 83-117.

Claudy, J. (1972). A comparison of five variable weighting procedures. *Educational and Psychological Measurement*, 32, 311-322.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7, 249–253

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological methods*, *8*(3), 243.

Colom, R., Martínez-Molina, A., Shih, P. C., & Santacreu, J. (2010). Intelligence, working memory, and multitasking performance. *Intelligence*, *38*(6), 543-551.

*Cowan, D. K., Barrett, L. E., & Wegner, T. G. (1990). *Air Force officer training school selection system validation* (No. AFHRL-TR-89-65). Brooks AFB, San Antonio, TX: Air Force Human Resources Laboratory.

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Pedraza, O., Mehta, K. M., Tang, Y., ... & Mungas, D. M. (2008). Composite scores for executive function items: demographic heterogeneity and relationships with quantitative magnetic resonance imaging. *Journal of the International Neuropsychological Society*, *14*(5), 746-759.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

*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. and Gibb G. D., (1986). *Development of a Computer-Based Naval Aviation Selection Test Battery. (No.* NAMRL 1319). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

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. L. (1996). Pilot selection batteries: Shortcomings and perspectives. *The international journal of aviation psychology*, *6*(2), 199-209.

Damos, D. L. (2011). *A summary of the technical pilot selection literature*. (No. AFCAPS-FR-2011-0009). Randolph AFB, TX: Air Force Personnel Center, Strategic Research and Assessment Branch.

*Darr, W. (2009). A Psychometric Examination of the Canadian Automated Pilot Selection System (CAPSS). *Ottawa, Canada: Director General Military Personnel Research & Analysis*.

DeGeest, D. S., & Schmidt, F. L. (2010). The impact of research synthesis methods on industrial–organizational psychology: The road from pessimism to optimism about cumulative knowledge. *Research Synthesis Methods*, *1*(3-4), 185-197.

*Delaney, H. D. (1992). Dichotic listening and psychomotor task performance as predictors of naval primary flight-training criteria. *The International Journal of Aviation Psychology*, *2*(2), 107-120.

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.

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.

*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.

Duval, S. J., & Tweedie, R. L. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95, 89–98.

*Emery, B. (2011). *Neurocognitive predictors of flight performance of successful solo flight students* (Doctoral Dissertation). Retrieved from ProQuest Dissertations Publishing. 3489209.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109), 629-634.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380.

*Endsley, M. R., & Bolstad, C. A. (1994). Individual differences in pilot situation awareness. *The International Journal of Aviation Psychology*, *4*(3), 241-264.

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. Personnel Psychology, 49, 275-306.

Federico, P. A. (1989). *Computer-based and paper-based measurement of recognition performance* (No. NPRDC-TR-89-07). San Diego, CA: Navy Personnel Research and Development Center.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychological methods*, *6*(2), 161.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, 10, 444–467.

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

*Forgues, S. (2014). *Aptitude Testing of Military Pilot Candidates* (Doctoral dissertation). Retrieved from https://qspace.library.queensu.ca/bitstream/handle/1974/12582/Forgues_Susan_L-201410_MED.pdf?sequence=1.

*Forsman, J. W. (2012). *The Creation and Validation of a Pilot Selection System for a Midwestern University Aviation Department* (Doctoral dissertation). Retrieved from https://cornerstone.lib.mnsu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1153&context=etds.

*Gibb, G. D. (1990). Initial validation of a computer-based secondary selection system for student naval aviators. *Military Psychology*, *2*(4), 205.

Gibbons, L. E., Carle, A. C., Mackin, R. S., Harvey, D., Mukherjee, S., Insel, P., ... & Alzheimer's Disease Neuroimaging Initiative. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain imaging and behavior*, *6*(4), 517-527.

Glass, G. V. (1977). Integrating findings: the meta-analysis of research. *Review of research in education*, *5*(1), 351-379.

Gray, W. R. (2010). USAF Test Pilot Selection for the Next Generation. In: U.S. Air Force, *Test Pilot School, AF Instruction 99-107* (ed) U.S. Air Force T&E Days, Nashville, Tennessee, Sept 2002.

*Gress, W., & Willkomm, B. (1996). Simulator-based test systems as a measure to improve the prognostic value of aircrew selection. *Selection and Training Advances in Aviation: AGARD Conference Proceedings 588* (pp. 15-1-15-4). Prague, Czech Republic: Advisory Group for Aerospace Research & Development.

*Griffin, G. R. (1998). Predicting naval aviator flight training performance using multiple regression and an artificial neural network. *The International Journal of Aviation Psychology*, *8*(2), 121-135.

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.

Gress, W., & Willkomm, B. (1996). Simulator based test systems as a measure to improve the prognostic value of aircrew selection. *Progrès réalisés en sélection et formation des personnels navigants*, *588*, 12.

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology*, *87*(2), 377.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*(2), 388.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods*, *3*(4), 486.

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

Heaton, R. K., Akshoomoff, N., Tulsky, D., Mungas, D., Weintraub, S., Dikmen, S., ... & Gershon, R. (2014). Reliability and validity of composite scores from the NIH Toolbox Cognition Battery in adults. *Journal of the International Neuropsychological Society*, *20*(6), 588-598.

*Herniman, D. D. (2013). *Investigating Predictors of Primary Flight Training in the Canadian Forces* (Master Thesis). Retrieved from https://curve.carleton.ca/system/files/etd/

3c775297-cdf3-4784-b9bb-1f7b9db20811/etd_pdf/9d94fb2d52cded864d75540edb8acefa/herniman-investigatingpredictorsofprimaryflighttraining.pdf.

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

Hirsh, H. R., Northrop, L. C., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology*, *39*(2), 399-420.

Hoelzle, J. B. (2008). Neuropsychological assessment and the Cattell–Horn– Carroll (CHC) cognitive abilities model. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *69(9B)*. (UMI No. AAI 3328214).

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

*Hörmann, H. J., Luo, X. L., & Hamburg, G. G. (1999). Development and validation of selection methods for Chinese student pilots. In R. S. Jensen, B. Cox, J. D. Callister, & R. Lavis (EDs), *Proceedings of the Tenth International Symposium on Aviation Psychology* (pp. 571-576). Columbus, OH: Ohio State University.

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

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of educational psychology*, *57*(5), 253.

Hornby, R. (2003). *A multiple test battery approach during the assessment of the auditory nervous system of patients with multiple sclerosis* (Doctoral dissertation). Retrieved from https://repository.up.ac.za/bitstream/handle/2263/26535/dissertation.pdf?sequence=1&isAllowed=y.

Horst, P. (1951a). The relationship between the validity of a single test and its contribution to the predictive efficiency of a test battery. *Psychometrika*, *16*(1), 57-66.

Horst, P. (1951b). Optimal test length for maximum battery validity. *Psychometrika*, *16*(2), 189-202.

Howse, W. R., & Damos, D. L. (2011). *A bibliographic database for the history of pilot training selection* (No. DAS-2011-02). Randolph AFB, TX: Air Force Personnel Center.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index?. *Psychological methods*, *11*(2), 193.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, *81*(5), 459.

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

Hunter, D. R., & Burke, E. F. (1992). *Meta-analysis of aircraft pilot selection measures* (No. ARI-RN-92-51). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

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 alternate predictors of job performance. *Psychological Bulletin*, 96, 72–98.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. International Journal of Selection and Assessment, 8, 275-292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

*Ingurgio, V., & Crawford, C. V. (2017). *Revalidation of the Selection Instrument for Flight Training*. (No. Research Report 2002). Fort Belvoir, VA: Army Research Institute for the Behavioral and Social Sciences.

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

Johnson, J. F., Barron, L. G., Rose, M. R., & Carretta, T. R. (2017). Validity of Spatial Ability Tests for Selection into STEM (Science, Technology, Engineering, and Math) Career Fields: The Example of Military Aviation. In *Visual-spatial Ability in STEM Education* (pp. 11-34). Springer, Cham.

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of applied psychology*, *80*(1), 94.

Johnson, W., Bouchard Jr, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence*, *32*(1), 95-107.

Johnson, W., te Nijenhuis, J., & Bouchard Jr, T. J. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence*, *36*(1), 81-95.

Johnston, N. (1996). Psychological testing and pilot licensing. *The International journal of aviation psychology*, *6*(2), 179-197.

*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.

*Keener, R. A. (2003). *Use of Multivariate Techniques to Validate and Improve the Current USAF Pilot Candidate Selection Model* (No. AFIT/GOR/ENS/03-13). Wright-Patterson AFB, OH: U.S. Air Force Institute of Technology, School of Engineering and Management.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-Analytic Reporting Standards). *Journal of Business and Psychology*, *28*(2), 123-143.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276-2284.

King, R. E., Retzlaff, P., Barto, E., Ree, M. J., Teachout, M. S., & Carretta, T. R. (2012). *Pilot Cognitive Functioning and Training Outcomes* (No. AFRL-SA-WP-TR-2012-0004). Wright-Patterson AFB, OH: U.S. Air Force School of Aerospace Medicine.

*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, 28-38*.

King, R. E., Barto, E., Ree, M. J., Teachout, M. S., & Retzlaff, P. (2011). *Compilation of pilot cognitive ability norms* (No. AFRL-SA-WP-TR-2012-0001). Wright-Patterson AFB, OH: U.S. Air Force School of Aerospace Medicine.

*Kokorian, A., & Pereira da Costa, M. (2008, October). *On the transportability of a computerized test battery for the selection of pilots*. Paper presented at the 50th International Military Testing Association, Amsterdam, Belgium. Retrieved from IMTA Database (http://www.imta.info/PastConferences/PastConferences.aspx).

Kokorian, A., Valsler, C., Tobar, J. C., Force, C. A., Ribeiro, R. B., & Force, P. A. (2004, October). Generalizability of the Criterion Validity For a Pilot Selection Battery. Paper presented at the 46th *International Military Testing Association Symposium*, Brussels. Retrieved from IMTA Database (http://www.imta.info/PastConferences/PastConferences.aspx)..

*Kokorian, A., Valsler, C., & Burke, E. (2003, December). *International validation of a computerized testing suite for pilot selection*. Paper presented at the 6th Australian Aviation Psychology Symposium, Victoria, Australia. Retrieved from http://www.pilapt.com/downloads/international_validation_of_a_computerised_testing_uite_for_pilot_selection.pdf.

*Kokorian, A., Valsler, C., Tobar, J. C., Force, C. A., Ribeiro, R. B., & Force, P. A. (2004, October). *Generalizability of the criterion validity for a pilot selection battery*. Paper presented at the 46th International Military Testing Association, Brussels. Retrieved from IMTA Database (http://www.imta.info/PastConferences/PastConferences.aspx).

*Kole, M. L. (2006). *Predictors of flight performance in novice student pilots* (Doctoral dissertation). Retrieved from ProQuest Dissertation database. (UMI No. 3247234).

Konig, C. J., Buhner, M., & Murling, G. (2005). Working memory, fluid intelligence, and attention are predictors of multitasking performance, but polychronicity and extraversion are not. *Human performance*, *18*(3), 243-266.

Koslowsky, M., & Sagie, A. (1993). On the efficacy of credibility intervals as indicators of moderator effects in meta-analytic research. *Journal of Organizational Behavior*, *14*(7), 695-699.

*Lance, C. E., Stewart, A. M., & Carretta, T. R. (1993). *Refinement of Scoring Procedures for the Basic Attributes Test (BAT) Battery*. Brooks AFB, San Antonio, TX: Human Resources Directorate, Manpower and Personnel Research Division.

*Lance, C. E., Stewart, A. M., & Carretta, T. R. (1996). On the treatment of outliers in cognitive and psychomotor test data. *Military Psychology*, *8*(1), 43.

*Lehenbauer, L. P. (2004). *An investigation of the construct-related and criterion-related validity of CogScreen-Aeromedical Edition* (Doctoral Dissertation). Retrieved from ProQuest Dissertation database. (UMI No. 3088158).

Levine, E. L., Spector, P. E., Menon, S., & Narayanan, L. (1996). Validity generalization for cognitive, psychomotor, and perceptual tests for craft jobs in the utility industry. *Human performance*, *9*(1), 1-22.

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology*, *102*(1), 43.

Lipsey, M. W., & Wilson, D. B. 2001. Practical meta-analysis. Thousand Oaks, CA: Sage.

Lopez, R. A., & Denton, T. L. (2011). *Aviation Selection Test Battery Component Predictiveness of Primary Flight Training Outcomes Among Diverse Groups*. Monterey, CA: Naval Postgraduate School.

Lochner, K., & Nienhaus, N. (2016). *The predictive power of assessment for pilot selection*. Technical report published by cut-e Group. Retrieved April 15, 2018 from: https://www.cute.com/fileadmin/user_upload/Assessing_in_Aviation/White_Paper_Pilot_5238.pdf.

Lubinski D, Dawis RV. (1992). Aptitudes, skills and proficiencies. In Dunnette MD, Hough LM (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, 2nd ed., pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.

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.

*Maciejczyk, J., Kossowski, J. and Kuzak, W. (1995, October). *Evaluating correlation between scores yielded by pilot candidates on flight simulator and combat - training planes during elementary air training*. Paper presented at the 37th International Military Testing Association 1995 Conference, Toronto. Retrieved from IMTA Database (http://www.imta.info/PastConferences/PastConferences.aspx).

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

*Martinussen, M., Torjussen, T., Storsve, O., & Hjerkinn, O. (2004). *Pilot selection in the Norwegian Air Force: From paper and pencil to computer-based assessment*. Paper presented at the 40th Applied Military Psychology Symposium, Oslo.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment*, *8*(3), 290-302.

*McFarland, M. R. (2017). *Student Pilot Aptitude as an Indicator of Success in a Part 141 Collegiate Flight Training Program* (Doctoral dissertation). Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=kent1492088859648498&disposition=inline.

McGrew, K. S. (2005). The Cattell-Horn-Carroll Theory of Cognitive Abilities: Past, Present, and Future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 136-181). New York, NY, US: Guilford Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37(1–10).

McGrew, K. S., & Evans, J. J. (2004). Internal and external factorial extensions to the Cattell–Horn–Carroll (CHC) theory of cognitive abilities: a review of factor analytic research since Carroll's Seminal 1993 Treatise. (Carroll Human Cognitive Abilities Project Research Report No. 2). Retrieved February 23, 2018, from the *Institute for Applied Psychometrics* Website: http://www.iapsych.com/ carrollproject.htm.

Morgan, B., D'Mello, S., Abbott, R., Radvansky, G., Haass, M., & Tamplin, A. (2013). Individual differences in multitasking ability and adaptability. *Human factors*, *55*(4), 776-788.

*Morrison, T. R. (1988). *Complex Visual Information Processing: A Test for Predicting Navy Primary Flight Training Success* (No. NAMRL-1338). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

*Morrison, T. L. (1991, November). Validation of Biographical Inventory: US Navy/Marine Corps Aviation Selection Test Battery. *Proceedings of the 27th Meeting of the Department of Defense Human Factors Engineering Technical Advisory Group*, San Antonio, Texas.

*Naval Aerospace Medical Institute (NAMI) (1991). Aerospace Psychological Qualifications. In: *U.S. NAVAL Flight Surgeon's Manual*. Retrieved January 13, 2018 from file:///C:/Users/ksm_1/Downloads/480200.pdf.

*Ness, G. (1997). *Pilot Candidate Selection Method (PCSM) evaluation*. Randolph AFB, TX: AETC Studies and Analysis Squadron.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*(3), 301.

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.

*O'hare, D. (1997). Cognitive ability determinants of elite pilot performance. *Human factors*, *39*(4), 540-552.

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

*Olson, R. (2002). *An Analysis of Student Progress in Beginning Flight Training: Performance Prediction, Performance Measurement, and Performance Improvement* (Doctoral Dissertation). Retrieved from https://scholarworks.wmich.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=2299&context=dissertations.

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, and 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., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, 609-626.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2017). Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management. *Human Resource Management Review*, *27*(1), 201-215.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of educational statistics*, *8*(2), 157-159.

Ostoin, S. D. (2007). *An assessment of the performance-based measurement battery (PBMB), the Navy's psychomotor supplement to the Aviation Selection Test Battery (ASTB)*. Monterey, CA: Naval Postgraduate School.

*Park, K. S., & Lee, S. W. (1992). A computer-aided aptitude test for predicting flight performance of trainees. *Human Factors*, *34*(2), 189-204.

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.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training criteria in clerical occupations. *Journal of Applied Psychology*, 65, 373–407.

*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.

*Rani, E. K., & Chaturvedula, S. (2009). Accident proneness of pilots in Indian Air Force: An empirical analysis through selection criteria. *Indian Journal of Aerospace Medicine*, *53*(1), 36-44.

*Ree, M. J. (2003a). Test of Basic Aviation Skills (TBAS)–Scoring the Tests and Compliance of Tests with the Standards of the American Psychological Association. *San Antonio, TX: Operational Technologies Corporation*.

*Ree, M. J. (2003b). *Test of Basic Aviation Skills (TBAS) incremental validity beyond Air Force Officer Qualifying Test (AFOQT) pilot composite for predicting pilot criteria*. San Antonio, TX: Operational Technologies Corporation.

Ree, M. J., & Carretta, T. R. (1998). Computerized testing in the United States Air Force. *International Journal of Selection and Assessment*, *6*(2), 82-89.

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

Ree, M., Carretta, T., & Earles, J. (1998). In top-down decisions, weighting variables does not matter: A consequence of Wilk's theorem. *Organizational Research Methods*, *1*, 407-420.

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.

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

Retzlaff, P. D., King, R. E., & Callister, J. D. (1995). *USAF Pilot Training Completion and Retention: A Ten Year Follow-Up on Psychological Testing* (No. AL/AO-TR-1995-0124). Brooks AFB, TX: Armstrong Laboratory, Aerospace Medicine Directorate.

*Reweti, S., Gilbey, A., & Jeffrey, L. (2017). Efficacy of Low-Cost PC-Based Aviation Training Devices. *Journal of Information Technology Education: Research*, *16*, 127-142.

*Roomsburg, J. D. (1990). *Utility as a Function of Selection Ration and Base Rate: An Empirical Investigation of Military Aviation Selection* (No. AFIT/CI/CIA-90-004D). Wright-Patterson AFB, OH: Air Force Inst of Tech.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*(1), 59-82.

Roth P, Bobko P, McFarland L, & Buster M. (2008). Work sample tests in personnel selection: A meta-analysis of black-white differences in overall and exercise scores. *Personnel Psychology, 61*, 637–661.

Roth, P. L., Bobko, P., & McFarland, L. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, *58*(4), 1009-1037.

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Edison, C. E., Jr., & Schmit, J. J. (2005). Personality saturation in structured interviews. *International Journal of Selection and Assessment*, 13, 261–273.

Russell, T. L., Reynolds, D. H., & Campbell, J. P. (1994). *Building a joint-service classification research roadmap: Individual differences measurement* (No. FR-PRD-93-24). Alexandria, VA: Human Resources Research Organization.

Schneider, W. J., & Flanagan, D. P. (2015). The relationship between theories of intelligence and intelligence tests. In *Handbook of intelligence* (pp. 317-340). Springer, New York, NY.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99-144). New York, NY, US: Guilford Press.

Schneider, W. J., Mayer, J. D., & Newman, D. A. (2016). Integrating hot and cool intelligences: Thinking broadly about broad abilities. *Journal of Intelligence*, *4*(1), 1.

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.

Schmidt, F. L. (2015). History and development of the Schmidt–Hunter meta-analysis methods. *Research synthesis methods*, *6*(3), 232-239.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage publications.

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.

Schmidt, F. L., & Le, H. (2004). *Software for Hunter–Schmidt meta- analysis methods*. Iowa City, IA: University of Iowa, Department of Management and Organizations.

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 97-128.

Siem, F. M. (1992). Predictive validity of an automated personality inventory for Air Force pilot selection. *The International Journal of Aviation Psychology*, *2*(4), 261-270.

Siem, F., Carretta, T., & Mercatante, T. (1988). Personality, attitudes, and pilot training performance: Preliminary analysis (Interim Report, Jan. 1983- Jan. 1987).

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology*, *72*(1), 3.

*Spinner, B. (1991). Predicting success in primary flying school from the Canadian Automated Pilot Selection System: Derivation and cross-validation. *The International Journal of Aviation Psychology*, *1*(2), 163-180.

*Stauffer, J., & Ree, M. J. (1996). Predicting with logistic or linear regression: Will it make a difference in who is selected for pilot training?. *The International Journal of Aviation Psychology*, *6*(3), 233-240.

Sterne, J. A., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 75–98). West Sussex, United Kingdom: Wiley.

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, *87*(1), 96.

*Street, D. R., Chapman, A. E., & Helton, K. T. (1993). *The future of naval aviation selection: Broad-spectrum computer-based testing*. Paper presented at the 35th conference of the International Military Testing Association, Williamsburg, VA.

*Street Jr, D. R., & Dolgin, D. L. (1994). *Computer-based psychomotor tests in optimal training track assignment of student naval aviators* (No. NAMRL-1391). Naval Aerospace Medical Research Lab Pensacola FL.

*Stricker, L. J. (2005). The Biographical Inventory in Naval Aviation Selection: Inside the Black Box. *Military Psychology*, *17*(1), 55.

*Taylor, J. L., O'hara, R. U., Mumenthale, M. S., Yesavage, J. A., Taylor, J. L., O'harar, M. U., & Yesavage, J. A. (2000). Relationship of CogScreen-AE to flight simulator performance and pilot age. *Aviat Space Environ Med*, *7*, 373-80.

*Teachout, M. S., Ree, M. J., Barto, E. L., Carretta, T. R., King, R. E., & Michaels, C. F. (2013). *Consistency of pilot trainee cognitive ability, personality, and training performance in undergraduate pilot training*. San Antonio, TX: University of The Incarnate Word.

Tett, R. P., Hundley, N. A., & Christiansen, N. D. (2017). Meta-analysis and the myth of generalizability. *Industrial and Organizational Psychology*, *10*(3), 421-456.

Thomas, W. (2009). Minimizing the loss of student pilots from voluntary attrition. *Air & Space Power Journal*, *23*(4), 44.

Thorndike, R. L. (1987). Stability of factor loadings. *Personality and Individual Differences*, *8* (4), 585 – 586.

Tirre, W. C. (1997). Steps toward an improved pilot selection battery. In R. F. Dillon (Ed.), *Handbook on testing* (pp.220-255). Westport, CT: Greenwood Press.

*Tolton, R. G. (2014). *Relationship of Individual Pilot Factors to Simulated Flight Performance*, (Master thesis). http://hdl.handle.net/2142/23737.

Tsang, P. S., & Shaner, T. L. (1998). Age, attention, expertise, and time-sharing performance. *Psychology and aging*, *13*(2), 323.

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.

*Van Benthem, K., & Herdman, C. M. (2016a). Cognitive Factors Mediate the Relation Between Age and Flight Path Maintenance in General Aviation. *Aviation Psychology and Applied Human Factors*, 6, 81–90.

*Van Benthem, K., & Herdman, C. M. (2016b, December). Peripheral Motion Contrast Threshold as a Predictor of Aviator Performance. In *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale* (Vol. 70, No. 4, pp. 422-423). Ottawa, Ontario, Canada: Canadian Psychological Assoc.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological review*, *113*(4), 842.

Van der Maas, H. L., Kan, K. J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence*, *2*(1), 12-15.

Vernon, P. A. (1989). The generality of g. *Personality and Individual Differences, 10*, 803 – 804

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261-293.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48.

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.

Weissmuller, J. J., & Damos, D. L. (2014). Improving the pilot selection system: Statistical approaches and selection processes. *The International Journal of Aviation Psychology*, *24*(2), 99-118.

*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.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75*, 315–321.

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

Williams. A. T. (2009). Minimizing the loss of student pilots from voluntary attrition. *Air & Space Power Journal*, *23*(4), 44.

*Wingestad, Tore (2005). *Selection of offshore rotary wing pilots: Do psychological tests predict simulator performance?*. The Norwegian Defense Leadership Institute. Retrieved January 10, 2018 from http://docplayer.net/49141789-Selection-of-offshore-rotary-wing-pilots.html.

Wong, A., Keller, K. M., Sims, C. S., McInnis, B., Haddad, A., Giglio, K., & Lim, N. (2012). *The Use of Standardized Scores in Officer Career Management and Selection*. National Defense Research Institute. Retrieved March 7, 2018 from Rand Cooperation https://www.rand.org/pubs/technical_reports/TR952.html.

*Woycheshin, D. (2001). *Analysis of Canadian Automated Pilot Selection System (CAPSS) results*. (Air Personnel Research Report 01/1). Ottawa, Ontario, Canada: Chief of Air Staff. Retrieved February 17, 2018 from http://cradpdf.drdc-rddc.gc.ca/PDFS/unc01/p506202.pdf.

*Woychesin, D. E. (2002). Validation of the Canadian Automated Pilot Selection System (CAPSS) against primary flying training results. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *34*(2), 84.

Zachary, R. A. (1990). Wechsler's intelligence scales: Theoretical and practical considerations. *Journal of Psychoeducational Assessment*, *8*(3), 276-289.

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

Appendix A. A summary of the extracted information from the collected studies

(1) Test batteries saturated with Acquired Knowledge

Author(s)/Year	Organization	Battery	Composite	Criterion	*N*	*r*	Size	Regularity	Context	Experience
Arth et al. (1990)	USAF	AFOQT	Pilot	Dichotomous	695	0.21	Large	Common	Training	Novice
Baisden (1992)	US Navy	ASTB	AQT	Continuous	13755	0.36	Small	Common	Training	Novice
Baisden (1992)	US Navy	ASTB	AQT	Continuous	421	0.30	Small	Common	Training	Novice
Barron et al. (2016)	USAF	AFOQT	Pilot	Dichotomous	3140	0.10	Large	Common	Training	Novice
Barron et al. (2016)	USAF	AFOQT	Pilot	Ordinal	1662	0.11	Large	Common	Training	Novice
Bartram & Baxter (1996)	Cathay Pacific Airways	Selection Battery	Aptitude and GMA	Dichotomous	29	0.23	Small	Uncommon	Training	Novice
Blower & Dolgin (1991)	US Navy	ASTB	AQT/FAR	Continuous	557	0.12	Small	Common	Training	Novice
Blower (1992)	US Navy	ASTB	FAR	Contingency	836	0.07	Small	Common	Training	Novice
Blower (1998)	US Navy	ASTB	AQT	Continuous	936	0.34	Small	Common	Training	Novice
Boyd et al. (2005)	USAF	MAB	Verbal IQ	Ordinal	2105	0.14	Large	Uncommon	Training	Novice
Carretta & Ree (1994)	USAF	AFOQT	Pilot	Ordinal	678	0.20	Large	Common	Training	Novice
Carretta & Ree (1995)	USAF	MAB	Verbal IQ	Continuous	7563	0.19	Large	Common	Training	Novice
Carretta (1987a)	USAF	AFOQT	Pilot	Continuous	151	0.09	Large	Common	Training	Novice
Carretta (1987b)	USAF	AFOQT	Pilot	Continuous	526	0.13	Large	Common	Training	Novice
Carretta (1988)	USAF	AFOQT	Pilot	Continuous	110	0.07	Large	Common	Training	Novice
Carretta (1992)	USAF	AFOQT	Pilot	Dichotomous	696	0.13	Large	Common	Training	Novice
Carretta (1997)	USAF	AFOQT	Pilot	Dichotomous	9239	0.16	Large	Common	Training	Novice
Carretta (1997)	USAF	AFOQT	Pilot	Dichotomous	237	0.14	Large	Common	Training	Novice
Carretta (2005)	USAF	AFOQT	Pilot	Continuous	994	0.22	Large	Common	Training	Novice
Carretta (2011)	USAF	AFOQT	Pilot	Continuous	776	0.27	Large	Common	Training	Novice
Cowan et al. (1990)	USAF	AFOQT	Pilot	Dichotomous	1124	0.10	Large	Common	Training	Novice
Damitz et al. (2003)	German Major airline	DLR	Performance Competence	Continuous	73	0.29	Large	Common	Training	Novice
Delaney (1992)	US Navy	ASTB	FAR	Continuous	480	0.27	Small	Common	Training	Novice

Author(s)/Year	Organization	Battery	Composite	Criterion	*N*	*r*	Size	Regularity	Context	Experience
Forgues (2014)	Canadian Force	Selection Battery	CFAT	Simulator	1007	0.21	Small	Common	Training	Novice
Forsman (2012)	University program	Proposed test Battery	Mixed abilities	Continuous	18	0.03	Large	Uncommon	Training	Novice
Gibb (1990)	US Navy	ASTB	FAR	Continuous	415	0.28	Small	Common	Training	Novice
Gress & Willkomm (1996)	German AF	Selection Battery	Cognitive abilities	Continuous	267	0.30	Small	Common	Training	Novice
Griffin (1998)	US Navy	ASTB	FAR	Continuous	434	0.25	Small	Common	Training	Novice
Herniman (2013)	Canadian Force	CFAT	CFAT score	Continuous	75	0.11	Small	Common	Training	Novice
Ingurgio & Crawford (2017)	US Army	SIFT	Cognitive	Continuous	463	0.30	Small	Common	Training	Novice
Johnston & Catano (2013)	Canadian Force	CFAT	CFAT	Continuous	319	0.06	Small	Common	Training	Novice
Keener (2003)	USAF	AFOQT	Pilot	Dichotomous	6498	0.18	Large	Common	Training	Novice
King, et al. (2013)	USAF	MAB	Verbal IQ	Ordinal	12924	0.12	Large	Uncommon	Training	Novice
King, et al. (2013)	USAF	MicroCog	Reasoning/Calculation	Ordinal	5582	0.12	Small	Uncommon	Training	Novice
Martinussen & Torjussen (1998)	Norwegian AF	Selection Battery	Verbal Ability	Dichotomous	159	0.19	Small	Common	Training	Novice
McFarland (2017)	University program	ACT	Overall composite score	Continuous	96	0.08	Small	Uncommon	Training	Novice
Morrison (1988)	US Navy	ASTB	FAR	Continuous	405	0.19	Small	Common	Training	Novice
Morrison (1991)	US Navy	ASTB	AQT	Dichotomous	702	0.08	Small	Common	Training	Novice
NAMI (1991)	US Navy	ASTB	AQT	Continuous	1425	0.06	Small	Common	Training	Novice
Olea & Ree (1994)	USAF	AFOQT	g index	Continuous	1867	0.18	Large	Uncommon	Training	Novice
Phillips et al. (2001)	US Navy	ASTB	FAR	Continuous	248	0.35	Small	Common	Training	Novice
Rani & Chaturvedula (2009)	Indian AF	Selection	Intelligence index	Contingency	418	0.12	Small	Common	Training	Novice
Ree (2003b)	USAF	AFOQT	Pilot	Continuous	322	0.31	Large	Common	Training	Novice
Roomsburg (1990)	USAF	AFOQT	Pilot	Dichotomous	996	0.22	Large	Common	Training	Novice
Roomsburg (1990)	USAF	AFOQT	Pilot	Dichotomous	1185	0.11	Large	Common	Training	Novice
Roomsburg (1990)	USAF	AFOQT	Pilot	Dichotomous	812	0.13	Large	Common	Training	Novice
Roomsburg (1990)	USAF	AFOQT	Pilot	Dichotomous	764	0.16	Large	Common	Training	Novice
Stauffer & Ree (1996)	USAF	AFOQT	Pilot	Dichotomous	1228	0.16	Large	Common	Training	Novice
Street & Dolgin (1994)	US Navy	ASTB	FAR	Ordinal	237	0.27	Small	Common	Training	Novice
Street et al. (1993)	US Navy	ASTB	FAR	Continuous	159	0.25	Small	Common	Training	Novice

Author(s)/Year	Organization	Battery	Composite	Criterion	*N*	*r*	Size	Regularity	Context	Experience
Stricker (2005)	US Navy	ASTB	FAR	Dichotomous	1415	0.10	Small	Common	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	10	0.23	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	77	0.19	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	149	0.04	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	289	0.12	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	302	-0.03	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	357	0.20	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	379	0.08	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	359	0.08	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	345	0.08	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	157	0.19	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	179	-0.01	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	166	0.08	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	233	0.16	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Verbal IQ	Continuous	295	0.29	Large	Uncommon	Training	Novice
Williams et al. (2000)	US Navy	ASTB	FAR	Continuous	1660	0.36	Small	Common	Training	Novice
Woycheshin (2001)	Canadian Force	CFAT	CFAT score	Simulator	627	0.12	Small	Common	Training	Novice
Zierke (2014)	European Major airline	DLR	Knowledge tests	Dichotomous	402	0.14	Small	Common	Training	Novice

(2) Test batteries saturated with Perceptual Processing

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Size	Regularity	Context	Experience
Adamson et al (2010)	General Aviation Pilots	CogScreen	Cognitive speed of processing	Simulator	51	0.56	Large	Uncommon	Exploratory	Experienced
Arendasy et al. (2007)	German Luftwaffe	Selection Battery	Cognitive tests	Simulator	99	0.39	Large	Common	Training	Novice
Arth et al. (1990)	USAF	AFOQT	Navigator	Dichotomous	695	0.10	Large	Uncommon	Training	Novice
Baker (1989)	USAF	AFOQT	Navigator	Continuous	275	0.21	Large	Uncommon	Training	Novice
Barron & Rose (2017)	USAF	SynWin	Single Tasking	Continuous	370	0.03	Small	Uncommon	Training	Novice
Blower & Dolgin (1991)	US Navy	ASTB	Manikin/Baddeley	Contingency	557	0.13	Small	Common	Training	Novice
Boyd et al. (2005)	USAF	MAB	Performance IQ	Ordinal	2105	0.14	Large	Uncommon	Training	Novice
Carretta & Ree (1994)	USAF	BAT	Information Processing composite	Ordinal	678	0.03	Small	Common	Training	Novice
Carretta (1988)	USAF	AFOQT	Navigator	Continuous	110	0.29	Large	Uncommon	Training	Novice
Carretta (1992)	USAF	AFOQT	Navigator	Ordinal	696	0.08	Large	Uncommon	Training	Novice
Carretta (1997)	USAF	AFOQT	Navigator	Dichotomous	9239	0.13	Large	Uncommon	Training	Novice
Carretta (1997)	USAF	AFOQT	Navigator	Dichotomous	237	0.17	Large	Uncommon	Training	Novice
Carretta, 2005)	USAF	AFOQT	Navigator	Continuous	994	0.21	Large	Uncommon	Training	Novice
Cowan et al. (1990)	USAF	AFOQT	Navigator	Dichotomous	1124	0.08	Large	Uncommon	Training	Novice
Emery (2011)	University program	CogScreen	Factor of General Speed/Warking Memory	Dichotomous	52	0.28	Large	Uncommon	Training	Novice
Endsley & Bolstad (1994)	Northrop	Experimental	Memory	Simulator	25	0.07	Small	Uncommon	Exploratory	Experienced
Forgues (2014)	Canadian Force	Selection Battery	Perceptual speed	Simulator	1007	0.09	Small	Common	Training	Novice
Herniman (2013)	Canadian Force	Critical Reasoning Battery	Total score	Continuous	75	0.07	Small	Common	Training	Novice
Keener (2003)	USAF	AFOQT	Navigator	Dichotomous	6498	0.13	Large	Uncommon	Training	Novice
King, et al. (2013)	USAF	MAB	Performance IQ	Ordinal	12924	0.12	Large	Uncommon	Training	Novice
King, et al. (2013)	USAF	MicroCog	General Cognitive Functioning score	Ordinal	5582	0.20	Large	Uncommon	Training	Novice
Kole (2006)	University program	CogScreen	General Speed/ Working Memory	Continuous	39	0.38	Large	Uncommon	Exploratory	Novice

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Size	Regularity	Context	Experience
Lehenbauer (2004)	Commercial Airline	CogScreen	Visual Associative Working Memory	Continuous	398	0.11	Small	Uncommon	Training	Novice
Martinussen & Torjussen (1998)	Norwegian AF	Selection Battery	Spatial Ability	Dichotomous	159	0.19	Small	Common	Training	Novice
Morrison (1988)	US Navy	ASTB	Complex Visual Information Processing	Continuous	406	0.11	Small	Common	Training	Novice
Olson (2002)	University program	CogScreen	General Speed/ Working Memory	Continuous	23	0.62	Large	Uncommon	Exploratory	Novice
Phillips et al. (2001)	US Navy	ASTB	Emergency Scenario Test	Continuous	248	0.16	Small	Common	Training	Novice
Taylor et al. (2000)	Private-Licensed pilot	CogScreen	General Speed/ Working Memory	Simulator	100	0.57	Large	Uncommon	Exploratory	Experienced
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	10	0.44	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	77	0.06	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	149	0.16	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	289	0.14	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	302	0.14	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	357	0.17	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	379	0.13	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	359	0.00	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	345	0.15	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	157	0.13	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	179	-0.01	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	166	0.04	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	233	0.05	Large	Uncommon	Training	Novice
Teachout et al. (2014)	USAF	MAB	Performance IQ	Continuous	295	0.09	Large	Uncommon	Training	Novice
Tham (1995)	University program	Proposed Battery	Visual attention battery	Simulator	24	0.47	Large	Uncommon	Exploratory	Novice
Tolton (2014)	general aviation pilots	CogScreen	General Speed/ Working Memory	Simulator	54	0.28	Large	Uncommon	Exploratory	Experienced
Van Benthem & Herdman (2016)	general aviation pilots	CogScreen	Factor of General Speed/Working Memory	Simulator	54	0.46	Large	Uncommon	Exploratory	Experienced
Wingestad (2005)	Major offshore helicopter	Selection Battery	Pilot Prognosis	Simulator	100	0.51	Large	Uncommon	Exploratory	Experienced

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Size	Regularity	Context	Experience
Zierke (2014)	European Major airline	DLR	Cognitive Abilities	Dichotomous	402	0.14	Large	Common	Training	Novice

(3) Test batteries saturated with Motor Abilities

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Size	Regularity	Context	Experience
Adamson et al. (2010)	General Aviation Pilots	CogScreen	Perceptual-Motor Speed	Simulator	51	0.28	Large	Uncommon	Exploratory	Experienced
Bartram & Baxter (1996)	Cathay Pacific Airways	MICROPAT	SYNVAL	Dichotomous	29	0.29	Large	Common	Training	Novice
Bartram & Dale (1991)	UK Army	MICROPAT	P-Score	Dichotomous	81	0.11	Large	Common	Training	Novice
Bartram (1987)	UK Army	MICROPAT	P-Score	Dichotomous	243	0.13	Large	Common	Training	Novice
Blower & Dolgin (1991)	US Navy	ASTB-PB	Horizontal Tracking/Absolute Difference	Contingency	557	0.19	Small	Common	Training	Novice
Burke et al. (1997)	British Army Air Corps	Selection Battery	Pilot Aptitude Index	Dichotomous	341	0.35	Small	Common	Training	Novice
Burke et al. (1997)	British Royal Air Force	Selection Battery	Pilot Aptitude Index	Dichotomous	849	0.20	Small	Common	Training	Novice
Burke et al. (1997)	Turkish Air Force	Selection Battery	Pilot Aptitude Index	Dichotomous	570	0.23	Small	Common	Training	Novice
Carretta & Ree (1994)	USAF	BAT	Psychomotor coordination	Ordinal	678	0.16	Small	Common	Training	Novice
Carretta (2005)	USAF	TBAS	5-Tests Composite	Continuous	994	0.10	Large	Common	Training	Novice
Carretta (2011)	USAF	TBAS	TBAS	Continuous	776	0.28	Large	Common	Training	Novice
Delaney (1992)	US Navy	ASTB-PB	PMT task (Stick/Rudder/Throttle)	Continuous	480	0.26	Small	Common	Training	Novice
Forgues (2014)	Canadian Force	Selection Battery	Psychomotor Ability	Simulator	1007	0.45	Small	Common	Training	Novice
Griffin (1998)	US Navy	ASTB-PB	PMT task (stick & Rudar)	Continuous	434	0.36	Small	Common	Training	Novice
Hörmann et al. (1999)	China Civil Aviation	DLR-Chinese	Psychomotor coordination	Continuous	125	0.27	Small	Common	Training	Novice
Hörmann et al. (1999)	China Civil Aviation	DLR-Chinese	Psychomotor coordination	Continuous	200	0.27	Small	Common	Training	Novice
Keener (2003)	USAF	BAT	9-Tests Composite	Dichotomous	6498	0.08	Large	Common	Training	Novice
Kokorian et al. (2003)	UK RAF	PILAPT	PILAPT	Dichotomous	165	0.55	Large	Common	Training	Novice
Kokorian et al. (2004)	Chile Military	PILAPT	PILAPT	Dichotomous	67	0.36	Large	Common	Training	Novice
Kokorian et al. (2004)	Italy Military	PILAPT	PILAPT	Dichotomous	90	0.50	Large	Common	Training	Novice

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Size	Regularity	Context	Experience
Kokorian et al. (2004)	Portugal Military	PILAPT	PILAPT	Dichotomous	117	0.27	Large	Common	Training	Novice
Kokorian et al. (2008)	UK Civilian	PILAPT	PILAPT	Dichotomous	76	0.40	Large	Common	Training	Novice
Kokorian et al (2008)	Brazilian Air Force	PILAPT	PILAPT	Dichotomous	224	0.45	Large	Common	Training	Novice
Kole (2006)	University program	CogScreen	Motor Coordination	Continuous	39	-0.05	Small	Uncommon	Exploratory	Novice
Lance et al. (1993)	USAF	BAT	10-Tests Composite	Dichotomous	2451	0.10	Large	Common	Training	Novice
Lance et al. (1996)	USAF	BAT	6-Tests Composite	Dichotomous	2147	0.09	Large	Common	Training	Novice
Lehenbauer (2004)	Commercial Airline	CogScreen	Factor of Motor Coordination	Continuous	398	0.08	Small	Uncommon	Training	Novice
Martinussen et al. (2004)	Norwegian Air Force	Selection Battery	Total Index	Continuous	99	0.20	Large	Common	Training	Novice
O'hare (1997)	New Zealand Soaring Pilot	WOMBAT	Perceptual-Motor Coordination	Contingency	26	0.55	Small	Uncommon	Exploratory	Experienced
Phillips et al. (2001)	US Navy	ASTB-PB	ATT/VTT	Continuous	310	0.27	Small	Common	Training	Novice
Rani & Chaturvedula (2009)	Indian AF	Selection Battery	Flying Aptitude Scores	Contingency	550	0.11	Large	Common	Training	Novice
Ree (2003a)	USAF	TBAS	4-tests Composite	Continuous	551	0.33	Small	Common	Training	Novice
Ree (2003b)	USAF	TBAS	Airplane Tracking/Horizontal Tracking	Continuous	322	0.22	Small	Common	Training	Novice
Stauffer & Ree (1996)	USAF	BAT	Psychomotor Coordination	Dichotomous	1228	0.11	Large	Common	Training	Novice
Street & Dolgin (1994)	US Navy	ASTB-PB	PMT task (stick/rudder/throttle)	Ordinal	237	0.33	Small	Common	Training	Novice
Surrador et al. (2013)	Portuguese Air Force	Selection Battery	Perceptual-motor Coordination	Continuous	43	0.37	Small	Common	Training	Novice
Taylor et al. (2000)	Private-Licensed pilot	CogScreen	Motor Coordination	Simulator	100	0.21	Small	Uncommon	Exploratory	Experienced
Wheeler & Ree (1997)	USAF	BAT	General psychomotor tracking ability	Continuous	833	0.22	Large	Common	Training	Novice
Zierke (2010)	German DLR	DLR	Psychomotor Abilities	Dichotomous	402	0.11	Small	Common	Training	Novice

(4) Test batteries saturated with Controlled Attention

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Regularity	Context	Experience
Barron & Rose (2017)	USAF	SynWin	Memorization/Math/Visual Monitoring	Continuous	370	0.23	Uncommon	Training	Novice
Bartram & Dale (1991)	UK Army	MICROPAT	Dual Task	Dichotomous	81	0.06	Common	Training	Novice
Blower & Dolgin (1991)	US Navy	ASTB-PB	PMT/DLT	Dichotomous	641	0.10	Common	Training	Novice
Carretta (1988)	USAF	AFOQT	Encoding Speed	Continuous	110	0.20	Common	Training	Novice
Carretta (2005)	USAF	TBAS	Dual Tracking/Listening	Continuous	994	0.09	Common	Training	Novice
Delaney (1992)	US Navy	ASTB-PB	PMT/DLT	Continuous	480	0.28	Common	Training	Novice
Emery (2001)	University program	CogScreen	Divided Attention	Dichotomous	52	0.29	Uncommon	Training	Novice
Gibb (1990)	US Navy	ASTB-PB	Dual Tracking	Continuous	373	0.20	Common	Training	Novice
Griffin (1998)	US Navy	ASTB-PB	Third PMT/DLT	Continuous	434	0.36	Common	Training	Novice
Hörmann et al. (1999)	China Civil Aviation	DLR-Chinese	Multiple Tasks	Continuous	125	0.23	Common	Training	Novice
Hörmann et al. (1999)	China Civil Aviation	DLR-Chinese	Multiple Tasks	Continuous	200	0.49	Common	Training	Novice
Keener (2003)	USAF	BAT	Time Sharing	Dichotomous	6498	0.08	Common	Training	Novice
King et al. (2013)	USAF	CogScreen	Tracking/Delayed Memory	Ordinal	7003	0.02	Uncommon	Training	Novice
Kokorian et al. (2008)	UK Civilian	PILAPT	Capacity Battery	Dichotomous	76	0.28	Common	Training	Novice
Kokorian et al. (2008)	Brazilian Air Force	PILAPT	Capacity Battery	Dichotomous	224	0.33	Common	Training	Novice
Kokorian et al. (2008)	Italy Military	PILAPT	Capacity Battery	Dichotomous	90	0.36	Common	Training	Novice
Kole (2006)	University program	CogScreen	Divided Attention	Continuous	39	0.02	Uncommon	Exploratory	Novice
Lance et al. (1993)	USAF	BAT	Scanning/Allocating	Dichotomous	946	0.11	Common	Training	Novice
Lehenbauer (2004)	Commercial Airline	CogScreen	Tracking Dual Task	Continuous	398	0.18	Common	Training	Novice
Park & Lee (1992)	Korian AF	Selection Battery	Tracking/Memory	Dichotomous	63	0.29	Common	Training	Novice
Phillips et al. (2001)	US Navy	ASTB-PB	PMT/DLT	Continuous	248	0.16	Common	Training	Novice
Ree (2003b)	USAF	BAT	Emergency Scenario	Continuous	322	0.13	Common	Training	Novice
Street & Dolgin (1994)	US Navy	ASTB-PB	DLT/PMT (S)	Ordinal	237	0.31	Common	Training	Novice

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Regularity	Context	Experience
Street et al. (1993)	US Navy	ASTB-PB	Dual PMT/DLT	Continuous	159	0.34	Common	Training	Novice
Surrador et al. (2013)	Portuguese AF	Selection Battery	Psychomotor/Audio/Visual	Continuous	43	0.39	Common	Training	Novice
Taylor et al. (2000)	Private-Licensed pilot	CogScreen	Dual Tracking	Simulator	100	0.39	Uncommon	Exploratory	Experienced
Tolton (2014)	general aviation pilots	CogScreen	Dual Tracking	Simulator	54	0.26	Uncommon	Exploratory	Experienced
Van Benthem & Herdman (2016)	general aviation pilots	CogScreen	Visual attention	Simulator	54	0.48	Uncommon	Exploratory	Experienced
Van Benthem & Herdman (2016)	Licensed pilots	CogScreen	Dual Task	Simulator	24	0.05	Uncommon	Exploratory	Experienced

(5) Test batteries saturated with General Ability

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Regularity	Context	Experience
Baker (1989)	USAF	PCSM	Pilot/BAT/flying exp.	Continuous	275	0.34	Regular	Training	Novice
Boyd et al. (2005)	USAF	MAB	Full scale IQ	Ordinal	2105	0.14	Irregular	Training	Novice
Carretta & Ree (1994)	USAF	AFOQT	All 16 subtests	Ordinal	678	0.24	Irregular	Training	Novice
Carretta (1990)	USAF	BAT & AFOQT	Exploratory Composite	Dichotomous	430	0.21	Regular	Training	Novice
Carretta (1990)	USAF	BAT & AFOQT	Exploratory Composite	Dichotomous	455	0.20	Regular	Training	Novice
Carretta (2000)	USAF	PCSM	Pilot/BAT/flying exp.	Dichotomous	1268	0.34	Regular	Training	Novice
Carretta (2011)	USAF	PCSM	Pilot/BAT/flying exp.	Continuous	776	0.29	Regular	Training	Novice
Duke & Ree (1996)	USAF	PCSM	Pilot/BAT/flying exp.	Continuous	1082	0.27	Regular	Training	Novice
Emery (2001)	University program	CogScreen	LRPV	Dichotomous	52	0.32	Irregular	Training	Novice
Ingurgio & Crawford (2017)	US Army	SIFT	Total score	Continuous	463	0.33	Regular	Training	Novice
Keener (2003)	USAF	PCSM	Pilot/BAT/flying exp.	Dichotomous	6498	0.19	Regular	Training	Novice
King et al. (2013)	USAF	MAB	Full scale IQ	Ordinal	12924	0.16	Irregular	Training	Novice
Lance et al. (1993)	USAF	BAT	Exploratory Composite	Dichotomous	2451	0.09	Regular	Training	Novice
Lehenbauer (2004)	Commercial Airline	CogScreen	LRPV	Continuous	398	0.09	Irregular	Training	Novice
Martinussen & Torjussen (1998)	Norwegian AF	Selection Battery	7-tests Composite	Dichotomous	159	0.09	Regular	Training	Novice
Ness (1997)	USAF	PCSM	Pilot/BAT/flying exp.	contingency	576	0.29	Regular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	10	0.39	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	77	0.15	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	149	0.13	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	289	0.15	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	302	0.06	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	357	0.22	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	379	0.13	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	359	0.04	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	345	0.01	Irregular	Training	Novice

Author(s)/Year	Service	Battery	Composite	Criterion	*N*	*r*	Regularity	Context	Experience
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	157	0.19	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	179	-0.02	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	166	0.07	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	233	0.13	Irregular	Training	Novice
Teachout et al. (2014)	USAF	MAB	Full scale IQ	Continuous	295	0.23	Irregular	Training	Novice
Zierke (2014)	European Major airline	DLR	11-tests Composite	Dichotomous	402	0.19	Regular	Training	Novice

(6) Test batteries of Work Sample

Author(s)/Year	Service	Battery	Criterion	*N*	*r*	Status	Context	Experience
Darr (2009)	Canadian Force	CPASS	Continuous	403	0.31	Common	Training	Novice
Gress & Willkomm (1996)	German AF	Simulator	Continuous	267	0.44	Common	Training	Novice
Herniman, 2013)	Canadian Force	CPASS	Continuous	75	0.15	Common	Training	Novice
Hoermann & Goerke (2014)	Lufthansa's Flight Training Center	Simulator	Dichotomous	88	0.29	Common	Training	Novice
Johnston & Catano (2013)	Canadian Force	CPASS	Continuous	319	0.06	Common	Training	Novice
Maciejczyk et al. (1995)	Polish Airforce	Simulator	Continuous	64	0.32	Common	Training	Novice
Reweti et al. (2017)	Different Civilian organizations	PC Simulator	Continuous	62	0.43	Uncommon	Exploratory	Experienced
Spinner (1991)	Canadian Force	CPASS	Dichotomous	223	0.76	Common	Training	Novice
Woychesin (2002)	Canadian Force	CPASS	Continuous	154	0.35	Common	Training	Novice