REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
22-08-2014	Journal Article	03 June 2013 – 19 June 2014

4. TITLE AND SUBTITLE

Consistency of the Relations of Cognitive Ability and Personality Traits to Pilot Training Performance

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)

Thomas R. Carretta[1], Mark S. Teachout[2], Malcolm James Ree[3], Erica L. Barto3, Raymond E. King[4], and Charles F. Michaels[5]

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
53290902

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
See back of this form.

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711 Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Supervisory Control and Cognition Branch
Wright-Patterson AFB OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHCI

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-WP-RH-TR-xxxx-xxxx

12. DISTRIBUTION / AVAILABILITY STATEMENT
Distriubution A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
88 ABW Cleared 08/22/2014; 88ABW-2014-3944.

14. ABSTRACT

The predictive validity of cognitive ability and personality traits was examined in large samples of US Air Force pilot trainees. Criterion data were collected between 1995 and 2008 from four training bases across three training tracks. Analyses also examined consistency in pilot aptitude and training outcomes. Results were consistent with previous research indicating cognitive ability is the best predictor of pilot training performance. There were few differences across training tracks, bases, and years and none were large. Overall, results illustrated the consistency of the quality of pilot trainees as assessed by cognitive ability and personality trait measures, and the consistency of these measures in predicting training performance over time. This consistency results in a more stable training system, enabling greater efficiency and effectiveness.

15. SUBJECT TERMS
Pilot training performance, pilot aptitude, cognitive ability, personality traits, MAB, NEO-PI-R

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Antonio Ayala
Unclassified	Unclassified	Unclassified	SAR	38	**19b. TELEPHONE NUMBER** *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

7. PERFORMING ORGANIZATIONS NAMES AND ADDRESSES:

Thomas R. Carretta[1]
Air Force Research Laboratory
Wright-Patterson AFB, OH

Mark S. Teachout[2]
H-E-B School of Business and Administration
University of the Incarnate Word
San Antonio, TX

Malcolm James Ree[3]
Operational Technologies, Corporation
San Antonio, TX

Erica L. Barto[3]
Operational Technologies, Corporation
San Antonio, TX

Raymond E. King[4]
Headquarters, Air Force Safety Center
Kirtland AFB, NM

Charles F. Michaels[5]
Air Education and Training Command
Randolph AFB, TX

Running Header: Consistency of pilot attributes

Consistency of the Relations of Cognitive Ability and Personality Traits to Pilot Training Performance

Thomas R. Carretta
Air Force Research Laboratory
Wright-Patterson AFB, OH

Mark S. Teachout
H-E-B School of Business and Administration
University of the Incarnate Word
San Antonio, TX

Malcolm James Ree
Erica L. Barto
Operational Technologies, Corporation
San Antonio, TX

Raymond E. King
Headquarters, Air Force Safety Center
Kirtland AFB, NM

Charles F. Michaels
Air Education and Training Command
Randolph AFB, TX

The opinions expressed are those of the authors and not necessarily those of the United States Government, Department of Defense, or the United States Air Force. Address all correspondence to Dr. Thomas R. Carretta, AFMC 711th HPW/RHCI, 2210 8th Street, Area B, Bldg 146, Room 122, Wright-Patterson AFB, OH 45433-7511, e-mail: Thomas.Carretta@us.af.mil.

Abstract

The predictive validity of cognitive ability and personality traits was examined in large samples of US Air Force pilot trainees. Criterion data were collected between 1995 and 2008 from four training bases across three training tracks. Analyses also examined consistency in pilot aptitude and training outcomes. Results were consistent with previous research indicating cognitive ability is the best predictor of pilot training performance. There were few differences across training tracks, bases, and years and none were large. Overall, results illustrated the consistency of the quality of pilot trainees as assessed by cognitive ability and personality trait measures, and the consistency of these measures in predicting training performance over time. This consistency results in a more stable training system, enabling greater efficiency and effectiveness.

Key Words: pilot training performance, pilot aptitude, cognitive ability, personality traits, MAB, NEO PI-R

Consistency of the Relations of Cognitive Ability and Personality to Pilot Training Performance

The selection and training of military pilots is paramount to the success of the pilots and the military mission. The selection of military pilot trainees is a vital and critical task. Not only are pilots highly valued, they are expensive to train. The dollar costs of training are high and the risk to life and property are great. Therefore, it is important to ensure that the quality of pilot candidates remains high and stable over time, permitting pilot training to be as efficient and effective as possible. This paper examines the predictive validity of cognitive ability and personality measures for US Air Force (USAF) pilot trainees and the consistency of these relations across training tracks, bases, and time.

Background

The training of USAF pilots takes place in phases and at several different locations. Some of these locations also train pilots for other military services, both US and international. For example, US Navy aviators and European or other international military train at USAF facilities. Pilot training consists of three phases – academic classes and pre-flight training, primary aircraft training, and advanced aircraft training. Academic and pre-flight training course content includes aerospace physiology, ejection seat/egress/parachute landing, aircraft systems, instruments, mission planning, navigation, and weather. Primary and advanced (fighter/bomber or airlift/tanker) aircraft training is designed to teach flying skills with a focus on combat, instruments, formation, and navigation. While each training location follows roughly the same training syllabus to ensure coverage of common knowledge, skills, and abilities required for success, there are differences, with the Euro-NATO Joint Jet Pilot Training (ENJJPT) program at Sheppard Air Force Base being the most divergent (King & Lochridge, 1991). The ENJJPT

program is focused on training of combat pilots. Unlike Specialized Undergraduate Pilot Training (SUPT) which is taught at Columbus, Laughlin, and Vance Air Force Bases, ENJJPT has no airlift/tanker advanced training track. Also, ENJJPT students receive more hands-on flying hours in both the Primary and Advanced T-38 phases than those attending SUPT (see http://www.baseops.net/militarypilot/). A more detailed description of Primary and Advanced training are provided in the Method section.

The primary purpose of this study was to examine the predictive validity of cognitive ability and personality across three training tracks and four training bases over a 14-year period. Determining the generalizability of the predictive validity of these constructs is important as they have been mainstays in pilot selection batteries for many years (Carretta & Ree, 2003). A secondary purpose was to examine the consistency of pilot trainee quality and training performance across training tracks, bases, and time period. Maintaining a consistently high level of pilot trainee quality and training performance over time is crucial to ensuring the stability and effectiveness of the Air Force. Consistency should mean fewer changes and costs due to changes. Pilot trainee quality was measured using standardized tests of cognitive ability and personality traits. Training performance was measured using a composite of flying grades developed by USAF Air Education and Training Command (AETC).

USAF Pilot Candidate Selection Methods

All USAF pilot training applicants must pass the rigorous Class I flight physical standards (United States Air Force, 2011) to be eligible for selection. Medically qualified applicants are evaluated for training suitability on measures of officership and aptitude (Weeks & Zelenski, 1998). USAF Academy cadets are evaluated by Academy faculty and staff who consider

academic, military, and physical performance. Applicants commissioned through the Reserve Officer Training Corps (ROTC) or Officer Training School (OTS) are administered the Air Force Officer Qualifying Test (AFOQT; Drasgow, Nye, Carretta, & Ree, 2010) and Test of Basic Aviation Skills (TBAS; Carretta, 2005). A measure of pilot training aptitude, the Pilot Candidate Selection Method (PCSM; Carretta, 2011) score, is created by combining the AFOQT Pilot composite, several TBAS subtest scores, and the total number of flying hours logged either as a student pilot or as pilot in command[1] in a regression-weighted equation. For ROTC, medically qualified pilot training applicants are ranked on an Order of Merit score based on the PCSM score, field training, physical fitness, college grade point average (GPA), and commander's ranking. OTS pilot training candidate selection uses the "whole person" concept. Each OTS pilot training board member independently reviews the information in applicants' folders and scores each applicant in three areas: experience/leadership, education/aptitude, and potential/adaptability. If the scores for an applicant are not consistent across board members they discuss their scoring rationale until a sufficient level of agreement has been reached. Regardless of commissioning source, a common theme in pilot trainee selection procedures is high intelligence, whether it involves acceptance into the USAF Academy, a high GPA, a high AFOQT score, or the impression a candidate makes on a selection board.

Medical Flight Screening

In addition to the pilot trainee selection procedures described above, all candidates must complete Medical Flight Screening (MFS; King & Flynn, 1995). The USAF MFS program screens pilot candidates prior to Specialized Undergraduate Pilot Training (SUPT). MFS includes ophthalmic and cardiac diagnostic procedures as well as several psychological tests (King, Barto, Ree, Teachout, 2011; King, Barto, Ree, Teachout, & Retzlaff, 2011), including

[1] These are the number of flying hours in a FAA logbook and do not include hours in a flight simulator.

measures of cognitive ability (Multidimensional Aptitude Battery [MAB; Jackson, 2003] and MicroCog) and personality (NEO Personality Inventory – Revised [NEO PI-R; Costa & McCrae, 1985] and Minnesota Multiphasic Personality Inventory-2 [MMPI-2-RF; Butcher, Graham, Ben-Porath, Dahlstrom, & Kaemmer, 2001) tests.

Cognitive Tests. The primary purpose of the cognitive tests is to archive cognitive functioning data for future use in ideographic assessments where an individual is compared to themselves rather than to a collection of norms from a large population. The objective is to develop an individual registry against which future testing might be compared. Test results are particularly important for pilots seeking a waiver for return-to-flying status following an illness or injury that may have resulted in cognitive impairment (Chappelle, Ree, Barto, Teachout, & Thompson, 2010). During an evaluation, performance on the cognitive tests is compared with baseline scores collected prior to pilot training to determine whether any changes have occurred. Individualized (pre/post) comparisons result in more reliable return-to-flying duties decisions as pilots typically are very high cognitive functioning, especially in comparison to general population norms, and may remain so even after an injury or neurological event (King, 2012).

In addition to their clinical use, a recent study demonstrated that scores from the MAB and MicroCog were useful in predicting performance on several pilot training performance criteria including graduation/elimination from initial jet training and course grades (King, Carretta, Retzlaff, Barto, Ree, & Teachout, 2013). These results were consistent with prior studies of the relations of cognitive ability to pilot training performance (Carretta & Ree, 2003; Ree & Carretta, 1996).

Personality Tests. The US Air Force does not use measures of personality for pilot training selection. Measures of personality based on the Big Five model[2] (Goldberg, 1981) are administered by the Aeromedical Consultation Service USAF School of Aerospace Medicine prior to entry into pilot training. As with the MAB, these pre-training measures provide a baseline in subsequent psychological assessments when pilots are being considered for return-to-flying duties after receiving a medically disqualifying diagnosis. Archived personality test scores can be compared to the pilot's current functioning when seeking a waiver to the medical standards (United States Air Force, 2011). The operational personality assessment tool is the Revised NEO Personality Inventory (NEO PI–R; Costa & McCrae, 1985*)*, a Big Five measure which provides domain scores on Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness.

In one of the earliest reported studies of the use of personality tests for flying personnel, Sells (1955) showed the utility of the personality constructs of "motivation to fly" and "expression of anxieties about flying." Siem (1992) demonstrated the predictive validity of the personality constructs of hostility (r = -.12), self-confidence (r = .13), and values flexibility (r = .12) versus training completion in a sample of 509 USAF student pilots. Training graduates scored higher on self-confidence and values flexibility and lower on hostility than did those who failed due to flying training deficiency.

Anesgart and Callister (2001) examined the relationships between the NEO PI-R Big Five domain scores and success in flying training in a high-wing, propeller-driven monoplane. They reported that Neuroticism, Extraversion, and Openness were related to self-elimination from the program. Boyd, Patterson, and Thompson (2005) reported statistically significant

[2] The Big Five personality traits are five broad domains or dimensions used to describe human personality. The domains are neuroticism (sometimes called emotional stability), extraversion, openness, agreeableness, and conscientiousness.

differences between the scores of pilots assigned to fly airlift/tankers versus those assigned to fly fighters for the NEO PI-R domains of Agreeableness and Conscientiousness. Fighter pilots had lower levels of Agreeableness and higher levels of Conscientiousness.

Meta-analyses (Campbell, Castaneda, & Pulos, 2010; Hunter & Burke, 1994; Martinussen, 1996) have reported modest correlations between measures of personality and pilot raining performance. Hunter and Burke (1994) reported a small correlation (r = .10) for personality as a predictor of flying training criteria. Martinussen (1996) reported a small correlation (r = .14) for personality with training completion (pass/fail). More recently, Campbell et al. (2010) performed a meta-analysis on 26 studies examining the effects of personality as a predictor of pilot training completion (pass/fail). Two higher-order personality domains, Neuroticism (r = -.15) and Extraversion (r = .13), and one lower-order facet of Neuroticism, Anxiety (r = -.11), were found to have an impact on training success. After correction for range restriction and reliability of the predictors, the correlations were -.25 for Neuroticism, .17 for Extraversion, and -.14 for Anxiety. The authors concluded that emotionally stable, extroverted individuals would be better able to undergo the stress of aviation training. Finally, Chidester, Helmreich, Gregorich, and Geis, (1991) examined the relations between personality and crew coordination training performance in two samples of military pilots. Three profiles were identified through cluster analysis of the personality scales Positive Instrumental/Expressive, Negative Instrumental, and Low Motivation. These clusters replicated across samples and predicted attitude change following crew coordination training.

Purposes

The purpose of this study was to examine the predictive validity of cognitive ability and personality traits for pilot training performance. We also examined the consistency of pilot

trainee cognitive ability, personality traits and training success across three training tracks, four training bases and over a 14-year period. Maintaining a consistently high level of pilot trainee quality and training performance over time is crucial to ensuring an effective operational pilot cadre. Details regarding the predictor and criterion measures are provided in the Method section. Because consistency is vital to training success, fewer statistical differences are evidence of greater consistency and stability of the training system. To begin, we examined whether there were mean score differences in the cognitive, personality, and criterion scores across the training tracks, bases, and time period. Further, we examined the predictive validity of the cognitive and personality scores for pilot training performance. Here, consistency of prediction across tracks, bases, and time is important, as well as consistency with previous studies relating cognitive ability and personality to pilot training performance.

Method

Participants

A sample of 9,641 individuals selected for pilot training was administered the MAB and the NEO PI–R prior to beginning the 53 week Specialized Undergraduate Pilot Training (SUPT) program. All participants were college graduates or were near completion of college at time of testing. Selection ratios for pilot training assignments vary from year to year as a function of the number of applicants and the number of training positions available for each commissioning source. Of the participants reporting demographics information (98.5%), all were under the age of 36 years, with a modal age of 22 years, mean age of 24 years, and standard deviation of 2.6 years. Most of the participants (93%) were men. Racial and ethnic distributions indicated that 91% were White, 2% were African American, 3% were Hispanic, and 4% were “other.” All

were tested at either the School of Aerospace Medicine at Brooks City-Base, TX or at the USAF Academy in Colorado Springs, CO.

Measures

Multidimensional Aptitude Battery (MAB). The MAB (Jackson, 2003) is a broad-based test of intellectual ability patterned after the Wechsler Adult Intelligence Scale – Revised (WAIS-R; Wechsler, 1981). The MAB has 10 subtests that are combined to produce three summary scores: verbal IQ (VIQ), performance IQ (PIQ), and full-scale IQ (FSIQ). Previous research has demonstrated that the full-scale IQ score for the MAB and WAIS-R are strongly correlated (r = .91; Conoley & Kramer, 1989) and that the MAB measures general mental ability in several age groups (Wallbrown, Carmin, & Barnett, 1988). The MAB requires less than 1.5 hours to administer and can be individually or group administered. The subtests each have a normative mean of 50 and standard deviation of 10. FSIQ, VIQ, and PIQ scores have a mean of 100 and a SD of 15 in the general population. MAB norms are based on a sampling of nine age groups that were diverse in terms of gender, ethnicity, and race and North American (Canada and United States) geographic location. Test-retest reliability for the IQ scores ranges from .94 to .98 (Jackson, 2003) for an average retest interval of 45 days.

Table 1 provides brief descriptions and reliability of the subtests and indicates the summary IQ scores to which they contribute. Internal consistency reliability of the MAB–II in a sample of 91 twenty year olds was estimated using KR-20 (Jackson, 2003). This age group was the most similar to our participants. Reliabilities of the IQ scores ranged from .97 to .98 and reliabilities of the subtests ranged from .80 to .96.

[Insert Table 1 about here]

NEO PI–R. The NEO PI-R (Costa & McCrae, 1985) was designed to measure the Big Five personality domains and the facets or traits that underlie each domain. The five domains are Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. Each domain consists of six subscales called facet scores. These domains and facets provide a comprehensive measurement of adult personality.

The NEO PI–R was developed with the goal of being a multipurpose personality inventory useful for predicting many criteria such as behaviors related to illness, career interests, psychological health, and styles of coping (Costa & McCrae, 1985). It contains 240 statements that require examinees to respond on a Likert-type scale, ranging from 1) strongly disagree to 5) strongly agree. Table 2 provides description of the five domain scales as well as their internal consistency reliabilities (coefficient alpha) in a sample of 1,539 men and women in a large organization. Reliability coefficients for the 30 facets are reported in the test manual and range from .56 to .81 (Costa & McCrae, 1985). For the current study, the normative sample for adults served as the normative reference and the test was administered and scored via computer (Costa & McCrae, 1985).

[Insert Table 2 about here]

Training Performance Criterion. SUPT consists of a Primary aircraft training phase and an Advanced aircraft training phase. Primary aircraft training (T-6) consists of about 90 hours of flight training instruction over 22 weeks. The purpose is to teach basic flying skills including contact, instruments, formation (2-ship), and navigation. At the end of this phase, students are assigned to advanced training in either the fighter/bomber or the airlift/tanker track. Advanced training track assignments are a function of student preferences, training performance,

instructor ratings, and aircraft availability. The fighter/bomber advanced training track (T-38) includes about 120 hours of flight instruction over 24 weeks designed to prepare students for follow-on fighter/bomber training assignments. The initial training focus is on contact, instruments, formation (2/4 ship), navigation, and low-level flight. The airlift/tanker advanced training track (T-1) has about 115 hours of flight instruction over 26 weeks. The purpose is to prepare students for assignments to multiengine jet and turboprop aircraft. The training focuses on transition, instruments, navigation, low-level, and formation. It should be noted that training at Sheppard AFB differs from that at the other three bases. Sheppard AFB hosts the Euro-NATO Joint Jet Pilot Training (ENJJPT) program which is focused on training of combat pilots. It has no airlift/tanker advanced training track. Also, ENJJPT students receive more flying hours in both Primary (125 hours over 26 weeks) and Advanced T-38 (135 hours over 26 weeks) training than those attending SUPT.

The C-Score is a standardized flying training performance criterion measure developed by Headquarters Air Education and Training Command (AETC) in order to provide compatibility and comparability of performance at all US Air Force pilot training bases. The C-Score was developed after it was determined that there were mean differences in the ratings and other measures of pilot training performance across bases. For example, a very high scoring pilot at Base A might be scored lower than a high scoring pilot at Base B, due to idiosyncratic rating behavior by an instructor and/or check ride raters. As a result, comparisons across bases from one pilot training class to another were uncertain.

To enable meaningful comparisons (base-to-base, class-to-class, year-to-year, and pilot-to-pilot), the C-Score is a percentile rank based on a two year moving average. This allows the C-Score to reflect the training performance of each pilot, relative to the previous two years of

training performance for all pilots. Using past pilot performance as a moving-baseline average produces more reliable, stable, and interpretable scores, while permitting distinctions between individual performances.

The C-Score uses daily flying grades and check flight grades weighted approximately 1 to 2 in favor of the check flights. Daily flying grades include instructor pilots' evaluations of a pilot trainee's performance on all flights other than check flights. Daily flying grades are a weighted average of all flying training procedures/maneuvers performed during a flight and are rated unsatisfactory, fair, good, and excellent. In addition to daily flights, during training, pilot trainees must pass a check flight for each course of instruction. As with daily flying grades, check flight grades are a weighted average of ratings of flying procedures/maneuvers, which may have values of unsatisfactory, fair, good, and excellent. Maneuver grade point values are weighted based on the importance of the maneuver.

The C-Score calculation is standardized against approximately 200 previous students at that particular base or two years of students, whichever is greater. The calculations for each class are based on a "moving average," as one class is added to the population the oldest class is eliminated from the population. The C-Score is calculated for each class. Students are ranked on their C-Score value and each student is given a C-Score percentile rank, a number between 0% and 100%. The C-Score and percentile rank for a student are only recorded when the student is part of the graduating class.

Analyses

Analyses were conducted by training track, base, and year. Three analyses were conducted for each of these sets. First, descriptive statistics were calculated for the MAB IQ

Scores, NEO PI–R domain scores, and C-Score percentile rank. Second, analyses (*t*-tests or one-way ANOVAs) were conducted to determine statistical differences in mean scores for each variable in each category. Third, correlational analyses were conducted to determine how well the MAB and NEO PI–R scores predicted C-Score percentile rank.

Three sets of correlations were examined: observed (uncorrected) correlations, correlations corrected for range restriction, and correlations corrected for both range restriction and reliability of the scores. The assumptions underlying range restriction correction are the same as two of the three assumptions underlying the computation of a Pearson product-moment correlation - linearity of form and homoscedasticity. If the assumptions are met to estimate the correlation coefficient, they also are met to compute the correction. Restriction of range generally causes statistical indexes to underestimate true values. The multivariate correction method (Lawley, 1943) was used for the MAB-II scores. The univariate Case II correction (Thorndike, 1949) was used for the NEO-PI-R scores due to a lack of sufficient data to apply the multivariate method. The normative sample of the MAB–II and NEO PI–R provided the means, standard deviations, and correlations used for the correction. The corrected means, standard deviations, and correlations are superior estimates of the population values compared to the uncorrected values. This method removes the bias from the uncorrected sample estimates.

The range-restriction corrected correlations were then corrected for reliability (Hunter & Schmidt, 2004) of the test scores and training criterion ($r_c = \frac{r_{xy}}{\sqrt{r_{xx^*}}\sqrt{r_{yy}}}$.). The correlations were corrected for the reliability of *both* the test score and criterion because we were interested in the theoretical constructs underlying the measures, not the specific measures themselves. This third set of correlations provides a theoretical estimate of the validities of the underlying constructs when perfectly reliable measures are available.

Sample sizes differ for each analysis and are noted below each table. All analyses used a one-tailed test. The analyses that involved year-to-year comparisons used a .01 Type I error rate due to the large number of comparisons. All other analyses used a .05 Type I error rate. It should be noted, that while the very large samples used in this study ensure sufficient statistical power, very small differences will be statistically significant yet may offer little practical predictive power. Although we report statistical significance, because of the large samples involved, we focus on effect size (*d*, *r*). Importantly, fewer statistical differences (small effect sizes) across training tracks, bases, and years are desirable, as this indicates greater stability and consistency in the measures.

Results and Discussion

The predictive validity of cognitive ability and personality was examined in large samples of US Air Force pilot trainees by training track and training location for a 14 year period. Consistency in pilot aptitude and training outcomes was also examined. Validity results were consistent with previous findings that cognitive ability is the best predictor of pilot training performance (Carretta & Ree, 2003; Ree & Carretta, 1996).

Analyses by Training Track

The first set of analyses was conducted by training track, Primary, Advanced T–38, and Advanced T–1. Data were collapsed across training bases and years for these analyses.

Means. Descriptive statistics are shown in Table 3. The MAB IQ scores for the student pilots were severely range restricted compared to the normative values where the means and SDs are 100 and 15. The IQs for each of the training groups were high at about 120 (about 1.33 SD

above the normative mean) and the variances of the scores were much less than the normative values. For the FSIQ score the variance for the trainees was about 18% of the normative value.

The mean score differences between those assigned to the fighter/bomber and airlift/tanker tracks were small (1.27 points for the FSIQ or .20 *d*). The finding of slightly higher cognitive ability scores for fighter/bomber trainees is consistent with the selection and assignment of pilots for advanced training and with prior studies (Boyd et al., 2005). Because the T–38 track leads to more preferred assignments in fighter and bomber aircraft, higher cognitive ability students tend to be assigned to this track.

The results were mixed for the NEO PI-R where trainees were above the normative mean score of 50 for Extraversion and Conscientiousness and below the normative mean for Neuroticism and Agreeableness. For example, pilots score lower on Agreeableness than the general population (King, Barto, Ree, & Teachout, 2011). The lower mean for Agreeableness for trainees assigned to the fighter/bomber (T–38) advanced training track (T-38 = 42.70, T-1 = 44.38, d = -0.15) was consistent with previous results on personality for the highly selected pilot population

Independent-groups *t*-tests were conducted on each of the nine variables to identify significant differences between the two advanced training tracks. Because the advanced tracks include the students from the primary track, no comparisons were made with the primary track. Results indicated that there were small, but statistically significant mean differences between the T-38 and T-1 advanced tracks for 4 of the 9 scores. Cohen (1988) characterizes standardized mean differences (*d*) of .2 as small, .5 as medium, and .8 or greater as large. All mean score differences between trainees in the T-38 and T-1 tracks were small. T-38 trainees scored higher on the MAB VIQ (d = .31) and FSIQ (d = .20) scores than did T-1 trainees. However, T-38

trainees scored lower on the NEO PI-R Agreeableness score (d = -0.16) and the C-Score (d = -0.17) than those in the T–1 track.

[Insert Table 3 about here]

Correlations. Table 4 summarizes the correlational analyses by training track. All of the MAB IQ correlations with the C-Score were statistically significant for each training phase. Eight of the 15 correlations between the NEO PI-R scores and the C-Score were statistically significant. Cohen (1988) characterizes correlations of .10 as small, .30 as medium, and .50 or greater as large. All of the observed correlations between the MAB-II and NEO PI-R with the C-Score criterion were small. Even after correction for range restriction and reliability only 6 of the 24 correlations with the C-Score exceeded .30. These were for the MAB-II scores and C-Scores for the T-6 and T-38 tracks. Overall, the magnitudes of the correlations were higher for cognitive ability (MAB) than for personality traits (NEO PI–R).

[Insert Table 4 about here]

The overall correlational results for training tracks indicated that cognitive ability was related to pilot training success for all three tracks, and these correlations were higher than those for the personality trait measures. Small differences in the magnitude of validities of the cognitive test scores by training track were observed with lower values for T-1 training. For example, after correction for range restriction and reliability of the measures, the MAB FSIQ score validities were .377 for Primary (T-6), .386 for Advanced fighter/bomber (T-38), and .258 for Advanced airlift/tanker (T-1) training. The reason for these differences is unknown; however, they may be due to differing rater accuracy among other factors.

Analyses by Base

The second set of analyses was conducted by base (Columbus, Laughlin, Sheppard,[3] and Vance), for Primary, Advanced T–38, and Advanced T–1 training. Due to space limitations, the tables summarizing these analyses cannot be presented here. Interested readers should consult Teachout, Ree, Barto, Carretta, King, and Michaels (2013).

Means. Descriptive statistics were calculated for each of the 9 variables. One–way analyses of variance were conducted to identify any statistically significant differences among the bases for Primary, T-38, and T-1 training.

Primary training. The sample sizes by base for Primary training ranged from 1,023 to 2,781. Results indicated that there were small (Cohen, 1988), but statistically significant mean score differences between bases for six variables. Sheppard AFB differed from the other bases with Primary trainees about 2 points higher on all three MAB IQ scores. The standardized mean difference (*d*) on the FSIQ score between Sheppard and the other bases ranged from 0.33 to 0.39. Further, trainees at Sheppard were significantly lower on Agreeableness (about 1 point or 0.10 *d*) and higher on Conscientiousness (about 3 points or 0.31 *d*) than trainees at the other bases. These results may be due to the selectiveness of the Euro-NATO Joint Jet Pilot Training (ENJJPT) program.

Advanced T-38 training. The sample sizes by base for Advanced T-38 training ranged from 650 to 1,006. There were small, but statistically significant mean differences among the bases. The differences were between Sheppard and one or more of the other bases, paralleling the results for Primary training. The MAB scores at Sheppard were higher than for the other bases.

[3] Sheppard AFB, which hosts the combat-oriented Euro-NATO Joint Jet Pilot Training (ENJJPT) program, does not have an Advanced T-1 training track.

Advanced T-1 training. The sample sizes by base for Advanced T-1 training ranged from 351 to 589. There is no T-1 track at Sheppard AFB. Results indicated that there were small, but statistically significant mean differences among bases for only the C-Score. The C-Score for Columbus was significantly lower than Laughlin (d = -0.28) and Vance (d = -0.30).

Correlations. The pattern of correlations between the MAB and NEO PI-R scores and C-Score by base was similar to those observed when the data were collapsed across bases (see Table 4).

Primary training. For each base, all three MAB IQ scores demonstrated small, but statistically significant relations to the C-Score. For example, the correlations between the MAB FSIQ and C-Score ranged from.380 to .428 after correction for range restriction and reliability. The relations between the NEO PI-R scores and the C-Score were weaker than those for the MAB. Only 7 of 20 correlations were statistically significant.

Advanced T-38 training. Validities of the test scores for predicting T-38 training performance were generally lower and less consistent than those for Primary training. The correlation between the MAB FSIQ and C-Scores ranged from .170 to .458 after correction for range restriction and reliability. As with Primary training, the correlations between the NEO PI-R scores and C-Score were weaker than those for the MAB with only 7 of 20 NEO PI-R/C-Score correlations being statistically significant. Three of the 7 statistically significant correlations were for Openness.

Advanced T-1 training. As with T-38 training, results for T-1 training were less consistent than those for Primary training. The correlations between the MAB FSIQ and C-Scores ranged from .175 to .406 after correction for range restriction and reliability. The MAB PIQ score was not related to training performance for T-1 training.

Overall, the magnitude of the correlations was higher for cognitive ability (MAB) compared to personality traits (NEO PI–R). Only 2 of the 15 correlations between the NEO PI-R scores and the C-Score were statistically significant. Both were for Conscientiousness at Laughlin (.057) and Vance (.317) after correction for range restriction and reliability.

The most consistent result for comparisons of trainee quality across training bases was that Sheppard AFB had higher quality pilot trainees based on higher cognitive ability scores and higher scores on Conscientiousness, a key personality trait predictive of success in all jobs (Barrick & Mount, 1991). These pilots also were lower on Agreeableness. Further examination of student assignment to different bases is warranted to understand these differences. We can only speculate as to the underlying cause of these relations. Sheppard AFB is where the combat-oriented ENJJPT program is located. There is no separate advanced training track for non-fighter pilots. As a result, it is likely that pilot candidates who are considered to have a high probability of becoming fighter-qualified are assigned to ENJJPT.

Analyses by Year

The third set of analyses was conducted by year (1995–2008) for each training phase. Due to space limitations, the tables summarizing these analyses cannot be presented here but are available elsewhere (Teachout, et al., 2013).

Means. A one–way analysis of variance was conducted on each of the 9 scores to determine statistically significant differences among the 14 years for each training phase. The numerous comparisons for these analyses (91 comparisons for each of 9 scores for each phase = 819 comparisons/phase) should be viewed with caution, due to the increased likelihood of Type I error, that is, finding significant differences by chance as the number of comparisons increases.

For this reason, a $p < .01$ level of significance was used for comparing these mean differences. Further, rather than reporting and interpreting all of the significant differences, we focused on data trends. As described below, most of the statistically significant mean score differences occurred for Primary training. It is likely that Primary training attrition and the Advanced training assignment process contributed to making the Advanced training groups less variable.

Primary training. Results indicated there were statistically significant differences for 8 of the 9 scores for Primary training. Overall, while there were some statistically significant differences (75/819 = 9.1%), the scores were very stable, indicating that the characteristics and quality of pilot trainees were consistent over time. Further, all of the effect sizes were small. The number of significant differences was largest for the MAB PIQ score (22/91 = 24.1%) and C-Score (13/91= 14.3%). See Table 5. Sixteen of the 22 significant differences for PIQ were for years 2001-2003, where the PIQ scores were lower than for other years. For the C-Score, the mean for 1997 was higher than that for 1999 and 2003-2006 and the mean for 2002 was higher than those for 2003-2006.

[Insert Table 5 about here]

Advanced T-38 training. The degree of consistency in mean scores was greater for Advanced training than for Primary training. Six of the 9 scores exhibited significant differences for T-38 training. Only 3% (25/891) of the comparisons reached statistical significance. As with Primary training, all of the effect sizes were small and most of the significant differences occurred for the C-Score (11) and MAB PIQ (7). For the C-Score, 10 of the 11 differences occurred for 2000-2001 which were lower than other years. The MAB PIQ scores for 2005-2006 were higher than those for 2000-2003.

***Advanced T-1 training*[4].** Only 2 of 9 scores showed statistically significant mean score differences across year of training. Only 7.4% of the comparisons (4/54) were statistically significant indicating a remarkable degree of consistency in scores for the T-1 trainees.

Correlations. The correlational results broken out by year of training were consistent with those reported earlier where the data were collapsed across years. Overall, the magnitude of the correlations with the C-Score were higher for cognitive ability (MAB) than for personality traits (NEO PI–R).

Primary training. Although there was some variability, the magnitude of the correlations between the MAB and NEO PI-R scores with the C-Score by years was consistent and mirrored the results summed across years. Overall, the magnitude of the correlations was higher for cognitive ability compared to personality traits.

Advanced T-38 training. Again, the results broken out by year were consistent with those accumulated across years of training. The magnitude of the correlations with the C-score was higher for cognitive ability than personality traits.

Advanced T-1 training. Consistent with previous analyses, overall, the magnitude of the correlations with the C-Score was higher for cognitive ability than for personality traits. Further, there was little variability by year.

Given the large number of year-to-year comparisons made, the number of statistically significant differences was extremely small (5.6% across training tracks). This result illustrates the consistency of pilot selection methods and standards and their effect on trainee quality (cognitive ability and personality traits) over time. With pilot trainee characteristics this stable, fewer disruptions and adjustments are needed, the training system is more stable, enabling greater efficiency and effectiveness.

[4] T-1 training began in 2005. Prior to 2005 a different aircraft was used in airlift/tanker training.

There were more year-to-year differences noted in the C-Score. One possible explanation is fluctuation in managed attrition rates as projected manpower needs are adjusted by pilot training managers. Another possible source of score fluctuation is variation in the application of scoring criteria due to turnover in instructor pilots. More research is needed to investigate variability in C-Scores over time.

Results for personality trait measures were consistent with meta-analytic studies regarding the predictiveness of commonly used selection methods for both pilot training (Hunter & Burke, 1994; Martinussen, 1996) and in the broader context of personnel selection (Schmidt & Hunter, 1998).

Conclusions and Recommendations

While the observed validities for cognitive ability were small (Cohen, 1988), 6 of 9 correlations between the cognitive test scores and training criterion (see Table 4) were in the moderate range ($.3 \leq r \leq .5$) after correction for range restriction and reliability. The observed and corrected validities for personality traits were small and were consistent with previous studies (Anesgart & Callister, 2001; Campbell, Castaneda, & Pulos, 2010; Hunter & Burke, 1994; Martinussen, 1995; Siem, 1992). There were few differences across training tracks, bases, and years and none were large. The relative strength of the validities for the cognitive and personality trait measures was consistent with meta-analytic studies regarding the predictiveness of commonly used selection methods for both pilot training (Hunter & Burke, 1994; Martinussen, 1996) and in the broader context of personnel selection (Schmidt & Hunter, 1998).

The role of cognitive ability in pilot training has been to facilitate the acquisition of pilot job knowledge and flying skills (Ree, Carretta, & Teachout, 1995). The acquisition of

knowledge and skill in early pilot training has been shown to facilitate further knowledge and skills acquisition in later training. Path and structural equation models (Ree et al., 1995) showed the direct and indirect effects of cognitive ability on the acquisition of pilot job knowledge and flying skills. These direct and indirect effects probably account for the smaller validity coefficients for cognitive ability in advanced training in the current study. Additional studies are needed to examine the role of personality traits in the acquisition of pilot job knowledge and flying skills.

Overall, these results convey two notable messages. First, consistent with prior studies, measures of cognitive ability and personality traits are important determinants of pilot training success. Second, the quality of USAF pilot trainees has been remarkably consistent across training tracks and training locations over a 14 year period. This is likely a function of the availability of sufficient numbers of high quality applicants to fill available training positions and consistency in selection and training methods. These two messages are important for improving pilot selection and for practical application by decision-makers involved in setting selection and training requirements, and evaluating pilot training applicant suitability.

Improving Selection

The corrected validities were in the moderate range suggesting that there is a substantial proportion of criterion variance remaining to be predicted. The total amount of criterion validity that can be predicted is limited by external influences that may not be predictable. Student performance varies in pilot training for several reasons, not all of which are related to ability or personality traits. Some students may have personal problems that interfere with training performance. Others may have strong support from family that fortifies their training

performance. These and other outside influences should not be expected to be predicted by either cognitive ability or personality traits (Ree & Carretta, 1999).

Despite these limitations that reduce the magnitude of predictive relationships with pilot training outcomes, current USAF selection and classification methods do not leverage measures of cognitive ability and personality traits in an optimal manner to predict the remaining criterion variance. Although cognitive ability is represented in USAF pilot trainee selection methods such as the AFOQT and PCSM, measures of personality traits are not. Also, neither measures of cognitive ability nor personality traits are considered when making advanced training assignments. To this end, we recommend that studies be conducted to examine the incremental validity of personality measures for USAF pilot training qualification when used in combination with the PCSM score and measures of pilot aptitude. Further they should be examined to determine their utility in improving advanced training assignments when used in combination with Preliminary training performance, instructor ratings, and student preferences. Finally, measures of psychomotor performance should be included, as should measures of aviation-job knowledge and flying experience (Carretta & Ree, 2003).

Having good predictors is necessary but not sufficient for an optimal selection system. The criteria must be free of contamination and deficiency. As with predictors, criterion measures should be evaluated for evidence of construct validity. The identification of good criteria is just as important as the identification of good predictors.

Practical Applications

The current study demonstrated that pilot trainee quality and training performance were consistent over training track, training location, and time. The high quality of pilot trainees as

assessed by cognitive ability and personality trait measures and the consistency of these measures in predicting training performance over time enables the consistent production of high quality pilots. This stability in the selection and training system has multiple benefits. Importantly, Air Force decision-makers can rely on this stability for making policy, setting selection and training standards, and for longer-term planning activities (e.g., pilot production requirements). In addition, in the military aviation training system, consistency in trainee quality helps stabilize training methods (e.g., course content, instructional approaches, time and resources required to train students to meet rigorous standards). This enables the organization to meet its production goals (i.e., number of graduates) more efficiently and effectively over time.

References

Anesgart, M. N., & Callister, J. D. (2001). *Predicting training success with the NEO PI–R: The use of logistic regression to determine the odds of completing a pilot screening program,* AFRL-HE-WP-TR-2001-0074. Wright-Patterson AFB, OH: Air Force Research Laboratory, Human Effectiveness Directorate.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Boyd, J. E., Patterson, J. C., & Thompson, B. T. (2005). Psychological test profile of USAF pilots before training vs. type aircraft flown. *Aviation, Space, and Environmental Medicine, 76,* 463–468.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Dahlstrom, W. C., & Kaemmer, B. (2001). *MMPI-1 (Minnesota Multiphasic Personality Inventory-2) manual for administration, scoring, and interpretation, revised edition.* Minneapolis, MN: NCS Pearson.

Campbell, J. S., Castaneda, M., & Pulos, S. (2010). Meta-analysis of personality assessments as predictors of military aviation training success. *International Journal of Aviation Psychology, 20,* 92–109.

Carretta, T. R. (2005). *Development and validation of the Test of Basic Aviation Skills (TBAS),* AFRL-HE-WPTR-2005-0172. Wright-Patterson AFB, OH: Air Force Research Laboratory, Human Effectiveness Directorate.

Carretta, T. R. (2011). Pilot Candidate Selection Method: Still an effective predictor of US Air Force pilot training performance. *Aviation Psychology and Applied Human Factors, 1,* 3-8.

Carretta, T. R., & Ree, M. J. (2003). Pilot selection methods. In P.S. Tsang & M.A. Vidulich

(Eds.), *Principles and practice of aviation psychology (*pp. 357-396). Mahwah, NJ: Erlbaum.

Chappelle, W., Ree, M. J., Barto, E. L., Teachout, M. S., & Thompson, W. T. (2010). *Joint use of the MAB-II and MicroCog for improvements in the clinical and neuropsychological screening and aeromedical waiver process of rated USAF pilots,* AFRL-SA-BR-TR-2010-0002. Brooks City Base, TX: U.S. Air Force School of Aerospace Medicine.

Chidester, T. R., Helmreich, R. L., Gregorich, S. E., & Geis, C. E. (1991). Pilot personality and crew coordination: Implications for training and selection. *International Journal of Aviation Psychology, 1*, 25-44.

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

Conoley, J. C., & Kramer, J. J. (1989). *The tenth mental measurements yearbook.* Lincoln, NE: The University of Nebraska Press.

Costa, P. T., & McCrae, R. R. (1985). *The NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Drasgow, F., Nye, C., 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,* 68-85.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology*. Vol. 2 (pp. 141-165). Beverly Hills, CA: Sage.

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

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis.* Thousand Oaks, CA: Sage.

Jackson, D. N. (2003). *Multidimensional Aptitude Battery-II Manual.* Port Huron, MI: SIGMA Assessment Systems, Inc.

King, R. E. (2012). Psychological testing for mental health screening, suitability determinations, and archival purposes to improve safety and reduce costs. In A. Droog & M. Heese (Eds.), *Proceedings of the 30th Conference of the European Association for Aviation Psychology* (pp. 51-56). Sardinia, Italy: European Association of Aviation Psychology.

King, R. E., Barto, E., Ree, M. J., & Teachout, M. S. (2011). *Compilation of pilot personality norms,* AFRL-SA-WP-TR-2011-0008. Wright-Patterson AFB, OH: U.S. Air Force School of Aerospace Medicine.

King, R. E., Barto, E., Ree, M. J., Teachout, M. S., & Retzlaff, P. D. (2011). *Compilation of pilot cognitive ability norms,* AFRL-SA-WP-TR-2012-0001. Wright-Patterson AFB, OH: U.S. Air Force School of Aerospace Medicine.

King, R. E., Carretta, T. R., Retzlaff, P. D., 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., & Flynn, C. F. (1995). Defining and measuring the right stuff: Neuropsychiatrically enhanced flight screening (N-EFS). *Aviation, Space, and Environmental Medicine, 66,* 951–956.

King, R. E., & Lochridge, G. K. (1991). Flight psychology at Sheppard Air Force Base. *Aviation, Space, and Environmental Medicine, 62,* 1185-1188.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh,* LXII-Part 1, 19-32.

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

Ree, M. J., & Carretta, T. R. (1996). Central role of *g* in military pilot selection. *International Journal of Aviation Psychology, 6*, 111-123.

Ree, M. J., & Carretta, T. R. (1999). Lack of ability is not always the problem. *Journal of Business and Psychology, 14*, 165-171.

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

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

Sells, S. B. (1955). Development of a personality test battery for psychiatric screening of flying personnel. *Journal of Aviation Medicine, 26*, 35–45.

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

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*, AFRL-RH-WP-TR-2013-0081. Wright-Patterson AFB, OH: Air Force Research Laboratory, Decision Making Division.

Thorndike, R. L. (1949). *Personnel selection.* NY: Wiley.

United States Air Force (2011). *Medical examinations and standards*, Air Force Instruction 48-123. Washington, DC: Department of the Air Force.

Wallbrown, F. H., Carmin, C. N., & Barnett, R. W. (1988). Investigating the construct validity of the Multidimensional Aptitude Battery. *Psychological Reports, 62*, 871–878.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scales – Revised (WAIS-R).* NY: The Psychological Corporation.

Weeks, J. L., & Zelenski, W. E. (1998). *Entry to USAF undergraduate flying training*, AFRL-HE-AZ-TR-1998-0077. Brooks AFB, TX: Air Force Research Laboratory, Warfighter Training Research Division.

Table 1. *MAB-II Subtest and Summary Score Descriptions and Internal Consistency Reliabilities*

Scale	Subtest	Description	Reliability
VIQ, FSIQ	Information	Assesses the extent to which an individual has acquired knowledge about diverse topics	.87
VIQ, FSIQ	Comprehension	Measures the ability to evaluate social behavior, identify behavior that is more socially acceptable, and provide reasons why certain social customs and laws are practiced	.88
VIQ, FSIQ	Arithmetic	Assesses reasoning and problem solving ability through the solution of numerical problems	.80
VIQ, FSIQ	Similarities	Assesses the ability to conceptualize properties of an object and to compare them to those of another object, identifying the most similar characteristic	.90
VIQ, FSIQ	Vocabulary	Measures the ability to identify word meaning	.88
PIQ, FSIQ	Digit Symbol	Assesses visual-motor activity in substituting symbols for digits	.95
PIQ, FSIQ	Picture Completion	Measures the ability to identify missing elements in a picture	.88
PIQ, FSIQ	Spatial	Assesses the ability to visualize abstract objects in different positions in two-dimensional space	.96
PIQ, FSIQ	Picture Arrangement	Assesses the ability to arrange a set of randomly ordered pictures into a meaningful sequence	.85
PIQ, FSIQ	Object Assembly	Measures the ability to identify a complete object from disassembled	.89

Note. Reliability was estimated through internal consistency using KR-20 (Jackson, 2003). VIQ = Verbal IQ; FSIQ = Full-Scale IQ,; PIQ = Performance IQ.

Table 2. *NEO-PI-R Domain Definitions and Internal Consistency Reliabilities*

Test	Definition	Reliability
Neuroticism (N)	The tendency to experience negative emotions (anger, sadness, fear) and be emotionally unstable	.92
Extraversion (E)	The enjoyment of social situations, excitement, and stimulation	.89
Openness to Experience (O)	A willingness to explore new ideas and values; desire for aesthetics	.87
Agreeableness (A)	The desire to sympathize with and help others	.86
Conscientiousness (C)	Seeking a high-level of organization and planning; the tendency to plan carefully and exercise self-discipline	.90

Note. Reliability was estimated through internal consistency using Coefficient alpha for a developmental sample of 1,539 respondents (Costa & McCrae, 1985).

Table 3. *Descriptive Statistics for Primary and Advanced Training Tracks*

Score	Primary		Advanced T-38		Advanced T-1		T-38 vs. T-1	
	Mean	SD	Mean	SD	Mean	SD	*d*	*t*
C-Score	0.52	0.29	0.49	0.29	0.54	0.29	-0.17	-5.56**
VIQ	119.03	6.57	120.18	6.31	118.19	6.35	0.31	10.15**
PIQ	119.41	8.17	120.62	7.90	120.27	7.75	0.04	1.43
FSIQ	120.58	6.50	121.83	6.29	120.56	6.17	0.20	6.55**
N	46.65	9.37	46.07	9.46	46.29	9.17	-0.02	-0.79
E	57.59	9.56	58.12	9.65	57.65	9.47	0.05	1.58
O	50.67	10.18	50.49	10.39	50.05	9.66	0.04	1.39
A	43.81	10.56	42.73	10.66	44.38	10.28	-0.15	-5.12**
C	54.73	10.17	55.49	10.03	55.60	9.86	-0.01	-0.32

Note: Primary N = 9,396, Advanced T-38 N = 3,295, Advanced T-1 N = 1,524. VIQ = Verbal IQ; PIQ = Performance IQ; FSIQ = Full-Scale IQ N = Neuroticism; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness.

*$p < .05$; **$p < .001$

Table 4. *Observed and Corrected Correlations of MAB-II IQ Scores and NEO-PI-R Domain Scores with C-Score Percentile Rank by Training Track*

	Primary			Advanced T-38			Advanced T-1		
Score	*r*	r_c	r_{fc}	*r*	r_c	r_{fc}	*r*	r_c	r_{fc}
VIQ	.092**	.245	.321	.095**	.247	.324	.102**	.198	.260
PIQ	.117**	.266	.348	.115**	.275	.361	.056*	.150	.196
FSIQ	.126**	.288	.377	.126**	.295	.386	.098*8	.197	.258
N	-.023*	-.040	-.054	.014	-.020	-.027	.020	-.140	-.188
E	.008	-.060	-.082	.038*	-.050	-,068	-.002	-.090	-.123
O	-.064**	.050	.069	-.067**	.070	.097	-.042*	.060	.083
A	-.019*	-.030	-.042	-.059**	-.060	-.083	.029	.030	.041
C	.031**	.000	.000	.043*	.020	.027	.107**	.070	.095

Note. Sample sizes were Primary N = 9,396, Advanced T-38 N = 3,295, Advanced T-1 N = 1,524. Correlations in the column labeled *r* were observed (uncorrected). Those in the column labeled r_c were corrected for range restriction and those in the column labeled *rfc* were corrected for range restriction and reliability of the scores. The MAB IQ scores were corrected using the multivariate method (Lawley, 1943), while the NEO domain scores were corrected using the univariate Case 2 (Thorndike, 1949) method. Correlations in the column labeled r_{fc} were corrected for both range restriction and reliability of the test score and criterion. VIQ = Verbal IQ; PIQ = Performance IQ; FSIQ = Full-Scale IQ N = Neuroticism; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness.

*$p < .05$; **$p \leq .001$

Table 5. *Number of Statistically Significant Mean Score Differences across Years*

	Training Phase		
Score	Primary	Advanced T-38	Advanced T-1
C-Score	13	11	3
VIQ	3	0	0
PIQ	22	7	0
FSIQ	5	0	0
Neuroticism	9	1	0
Extraversion	0	2	0
Openness	2	0	0
Agreeableness	11	3	1
Conscientiousness	10	3	0
TOTAL	75	27	4

Note. The numbers indicate the number of statistically significant mean score differences at the $p < .01$ level. VIQ = Verbal IQ; PIQ = Performance IQ; FSIQ = Full-Scale IQ.