Predicting Commercial Pilot Training Performance: A Validation Study

Monica Martinussen[a,b], Ole Christian Lang-Ree[c], Håvard Mjøen[d], Bengt Svendsen[d], and Adrian Barone[d]

[a]Regional Centre for Child and Youth Mental Health and Child Welfare (RKBU Nord), Faculty of Health Sciences, UiT The Arctic University of Norway

[b]The Norwegian Defence University College, Oslo, Norway

[c]Norwegian Armed Forces Joint Medical Services, Oslo, Norway

[d]School of Aviation, Department of Engineering and Safety, UiT The Arctic University of Norway

Author Note

Correspondence concerning this article should be addressed to Monica Martinussen, RKBU-North, UiT The Arctic University of Norway, 9037 Tromsø, Norway. Phone: +4790133164. E-mail: monica.martinussen@uit.no

Declaration of Interest: All of the authors have been actively involved in the selection of individuals to the bachelor program in aviation for the UiT. The authors alone are responsible for the content and writing of the article. The views expressed in this article are those of the authors and are not necessarily those of the UiT or the Norwegian Air Force.

Acknowledgements: The study was financed by UiT The Arctic University of Norway.

Abstract

The main purpose of this study was to examine the predictive validity of the system used for ab initio selection of candidates to a bachelor program in aviation. The selection includes paper-and-pencil tests, computer-based tests, and an interview. One hundred eighty-eight candidates participated in the validation study. Total Test Score predicted the results of three exams in Aviation Theory (r = .27 – .38) and Extra Flying Hours Needed (-.22), but not Mean University Grade. The regression analyses indicated that all predictors (tests and interview ratings) explained 25% of the variance in Aviation Theory, 19% in Extra Flying Hours Needed, and 7% in Mean University Grade. The overall findings confirmed the predictive validity of selection tests, especially the computer-based tests.

Keywords: Selection, predictive validity, commercial pilots, testing, Big-five

Predicting Commercial Pilot Training Performance: A Validation Study

Personnel selection should utilize the best available methods and be conducted according to professional standards (AERA, 2014; EFPA, 2013; IATA, 2019). The term evidence-based selection implies that, when developing a selection system, evidence of the predictive validity of the tests employed in that system should be considered including a systematic analysis of job requirements (Broach, Schroeder, & Gildea, 2019), as should the expertise of those performing the selection, the needs of the applicant, and the needs of the organization (Martinussen, 2016). Moreover, revised European Union regulation 2018/1042 states that: “the operator shall ensure that flight crew has undergone a psychological assessment before commencing line flying” (EASA, 2018a). According to Annex III (EASA, 2018b), this should include an assessment of cognitive abilities, personality traits, operational and professional competencies, as well as social competencies in accordance with CRM principles. The Arctic University of Norway (UiT) has trained commercial pilots for more than 10 years, and this study represents the first evaluation of the predictive validity of the selection system.

Predicting Pilot Performance

Pilot selection in aviation has a long history dating back to World War 1, when the first selection tests were developed (Carretta & King, 2020). Several literature reviews (Carretta & Ree, 2003; Damos, 2011; Hunter, 1989; Martinussen, 2016; Paullin, Katz, Bruskiewicz, Houston, & Damos, 2006) and two large-scale meta-analyses (Hunter & Burke, 1994; Martinussen, 1996) have summarized findings on the predictive validity of selection tests across a large number of studies. The two meta-analyses were based on partly overlapping samples of studies and had consistent findings, showing that the test categories with the highest predictive validity were specific work sample tests and combined indices, followed by cognitive tests, psychomotor tests, and tests of aviation information (Hunter & Burke, 1994; Martinussen, 1996). The lowest predictive validity coefficients were found for personality measures, age, general intelligence tests, and academic grades. The mean predictive validity coefficients (uncorrected) ranged from .13 to .31 (Martinussen, 1996) and from -.10 to .34 (Hunter & Burke, 1994) for the different test categories.

Two more recent meta-analyses had more limited scopes: one summarized the predictive validity of multiple test batteries (ALMamari & Traynor, 2019), and the other looked at one specific test battery used for military selection in the United States (the Air Force Officer Qualification Test, AFOQT) (ALMamari & Traynor, 2020). The meta-analysis of multiple test batteries by ALMamari and Traynor (2019) was based on 118 independent samples consisting of either military (93 studies) or civilian pilots including university programs (25 studies), and found mean predictive validities (uncorrected) for the five different test categories ranging from .10 (Controlled Attention) to .34 (Work Sample). The meta-analysis of the AFOQT included a total of 32 samples from 26 studies and investigated the predictive validity of single measures and of the Total Pilot Score for this specific test battery. The mean (uncorrected) correlations with pilot training were .17 for the Pilot Composite and ranged between .00 (Word Knowledge) to .17 (Instrument Comprehension and Scale Reading) (ALMamari & Traynor, 2020) for single measures. An earlier meta-analysis of the different constructs measured by the AFOQT indicated a mean predictive validity of .15 for verbal ability and .24 for perceptual speed based on five studies (Johnson, Barron, Carretta, & Rose, 2017).

The Norwegian Air Force has used both paper-and-pencil tests as well as computer-based tests for selection of ab-initio candidates to the Norwegian Air Force (Martinussen & Torjussen, 1998; 2004). The computer-based test battery first introduced in 1998, has included between 11 and 13 tests, and two studies have examined their predictive validity (Martinussen & Torjussen, 2004; Lang-Ree, Martinussen, & Ødegaard, 2010). The (uncorrected) validities for the total test score was .26 (*N* = 108), and ranged between -.03 and .22 for single tests (Martinussen & Torjussen, 2004). The other study reported findings (uncorrected) for the total test score (.15) as well as for the three test categories .18 (Spatial Abilities), .08 (Psychomotor Coordination). and .06 (Information Processing) (*N* = 207) against pass/fail in training (Lang-Ree et al., 2010).

The majority of validation studies conducted so far have been based on military samples using training performance or pass/fail in training as a criterion. Only a small number of studies have provided evidence on the predictive value of a given test in the selection of commercial pilots. A study by the German Aerospace Center (DLR) included approximately 400 applicants to Lufthansa and indicated that the correlations between 11 cognitive ability and knowledge tests and pass/fail in training varied, some significantly, from .00 –.14 (uncorrected). The best predictor was Technical Comprehension (.14), followed Mathematics, Physics and one Spatial Orientation measure (.12). The multiple *R* was .19 (uncorrected) and .55 (corrected for range restriction and dichotomization) (Zierke, 2014).

Based on both meta-analyses and primary studies, there is strong evidence that cognitive and psychomotor tests predict future pilot performance, however some variation between studies have been detected (ALMamari & Traynor, 2019; 2020; Hunter & Burke, 1994; Martinussen, 1996).

Personality Traits and Other Characteristics as Predictors of Pilot Performance

The first meta-analyses indicated that personality tests in general had relatively low and variable validities when predicting pilot performance (Hunter & Burke, 1994;

Martinussen, 1996). However, after the introduction of the Big-five model, tests assessing these personality traits were examined for pilot selection. A small-scale meta-analysis utilizing the Big-five traits indicated mean uncorrected validity coefficients of -.15 for Neuroticism and .13 for Extroversion for predicting military training success (Campbell, Castaneda, & Pulos, 2009). Another study from the United States indicated relatively low predictive validities for the Big-five traits, based on a study of 883 students attending undergraduate pilot training who completed an experimental personality measure based on the Big-five model, in addition to two scales labelled Service Orientation and Team Orientation (Carretta, 2011). The findings indicated a small, incremental validity for the trait of Openness when predicting pass/fail, after controlling for the other predictors. Similarly, the traits Extraversion, Service Orientation, and Openness added a small incremental validity to some of the continuous training criteria (Carretta, 2011). Another study of 9,396 candidates who completed the NEO-PI-R prior to undergraduate pilot training (Carretta, Teachout, Ree, Barto, King, & Michaels, 2014) showed very low mean (uncorrected) correlations between Big-five traits and primary training results, ranging from -.06 for Openness to .03 for Conscientiousness (Carretta, et al., 2014).

So far, few studies have examined the use of personality traits in the selection of civilian pilots. The DLR developed trait-based measures in addition to using an assessment center and work sample tests to examine social, personal, and team competencies (Eißfeldt, 2014). An early validation study of the Temperament Structure Scales (TSS) supported the validity of four of those scales to predict job success as a pilot (*N* = 193-274) (Hörmann & Maschke, 1993; 1996). The authors concluded that successful pilots were characterized by higher values on the interpersonal scales and lower on emotional scales (Hörmann & Maschke, 1996), with correlations that varied between .17 and .29 in absolute values for specific traits and average check results (Hörmann & Maschke, 1993). A more recent study (Hörmann & Goerke, 2014) showed support for only a minority of the scales from two measures assessing social skills (SSI) and personality traits (TSS) when predicting training success (*N* = 88). Overall, the findings related to the predictive validity of personality measures for pilot selection are mixed and inconsistent.

Selection of Students to a Bachelor Program in Aviation

The bachelor program in aviation at the UiT includes both regular university courses (mathematics, physics, leadership and organizational studies, and a bachelor thesis) and the specific courses and training necessary to become a commercial pilot. The program accepts 12 new students every semester. The formal requirements for acceptance include completion of high school with a higher level in mathematics and physics.

The Current Study

The main purpose of this study was to examine the predictive validity of the selection system used for ab initio selection of candidates to the bachelor program in aviation. We examined the predictive validity of groups of selection tests and the selection interview as individual predictors, in addition to the predictive validity of all these predictors combined. We expected that the paper-and-pencil General Mental Ability (GMA) tests would predict pilot performance, especially GMA tests related to academic performance criteria, in accordance with previous findings (Martinussen, 1996). In addition, we expected that the computer-based tests assessing Psychomotor Coordination, Spatial Ability, and Information Processing would predict pilot performance, in line with previous meta-analyses (ALMamari & Traynor, 2019; Hunter & Burke, 1994; Martinussen, 1996) and earlier validation studies (Carretta et al., 2014; Martinussen & Torjussen, 2004; Lang-Ree, Martinussen & Ødegaard, 2010). We expected small correlations between Big-five Traits and the performance criteria, in line with meta-analyses and single studies (Campbell, et al., 2009; Carretta et al., 2014). Finally, we expected that the interview ratings would predict both academic and flying performance criteria as suggested in meta-analyses of predictors of work performance in general (Levashina et al., 2014; Schmidt et al., 2016)

Method

Participants and Procedure

The present analysis included 188 candidates who were accepted to the bachelor program in aviation over a 10-year time period. All candidates were between 19 and 44 years old (*M* = 22.83, *SD* = 3.51) when selected, and 12% were women. The project was approved by the Norwegian Center for Research Data. The mean number of applicants per semester was 131 (range between 80-161), resulting in an average selection ratio of 9% (12 students per semester/131 applicants × 100). After the initial screening and ranking of students based on high school grades, the best candidates (approximately 60 students or 46%) are invited to the next step of the selection process which includes testing and interview. Every step requires that candidate pass a certain minimum performance on the tests in order to proceed to the next phase. Approximately 60% of this group pass all the tests and are interviewed. The testing is conducted by the Norwegian Air Force Selection Center, and is a stepwise process where paper-and pencil tests are administered first, then computer-based tests and finally interview. The selection is usually conducted over two consecutive days and the final acceptance to the program is provided after the candidates have passed the medical examination conducted by the Institute of Aviation Medicine.

.

Predictors

The predictors included groups of paper-and-pencil and computer-based tests, personality traits, and interview ratings. A more detailed description of the tests and scoring may be found in previous publications (Martinussen & Torjussen, 1998; 2004). The total time for administering the paper-and-pencil tests including the personality measure is approximately three hours. In addition, the computer-based tests include 40 minutes for the psychomotor tests, 90 minutes for spatial abilities, and one hour for information processing. A small study has been conducted to examine the test-retest reliability of the four groups of tests based on a sub-sample of 51 applicants that had been re-tested after approximately 18 months (Lang-Ree, 2021). The findings were for the paper-and-pencil tests ($r_{tt} = .85$), psychomotor tests ($r_{tt} = .67$), spatial abilities ($r_{tt} = .81$), and information procession ($r_{tt} = .60$). Corresponding analyses were conducted for the total test score ($r_{tt} = .77$), and for the total computer-based test score ($r_{tt} = .69$) (Lang-Ree, 2021).

Paper-and-Pencil GMA Tests

General Mental Ability. Five measures were summarized in the overall mean GMA score: (a) *Raven Advanced Progressive Matrices* (Raven, 1986) as a measure of non-verbal abstract and analytical intelligence; (b) *Word Comprehension* as a measure of verbal intelligence; (c) *Number Series* as a measure of mathematical reasoning, in which the task is to fill in the two last digits in series of numbers; (d) *Mathematics* as a measure of mathematical knowledge and skills; and (e) *English Comprehension Test*.

Computer-Based Cognitive and Psychomotor Tests

Of the 11 computer-based tests (Martinussen & Torjussen, 2004), 10 were programmed by Psytech LTD (Burke, Kitching, & Valsler, 1997). The remaining computer-based test (Determinationsgerät, DTG) is part of the Vienna Test System developed by the Schuhfried Company (https://www.schuhfried.com/). The computer-based tests were categorized into three groups (Psychomotor Coordination, Spatial Ability, and Information Processing) based on a previous factor analysis (Martinussen & Torjussen, 2004). A mean score (Stanine) was computed for each group of tests, and a total computer-based test score was calculated based on all 11 computer-based tests.

Psychomotor Coordination. Three tests were included in this group: (a) *Control of Velocity (CVT-600),* a 1-dimentional pursuit tracking task; (b) *Sensory Motor Apparatus (SMA-610),* a two-dimensional compensatory tracking task; and (c) *Trax (630),* a pursuit tracking task modeled on the idea of flying down through an ILS glide path (PILAPT Handbook, 2006). The *Trax (630)* assesses a combination of psychomotor co-ordination, and spatial and information processing abilities.

Spatial Ability. This group included three tests: (a) *Planes-640,* a spatial ability test involving visualization and mental rotation; (b) *Instrument Comprehension (INSB-650),* a measure of general reasoning and spatial reasoning; and (c) *Hands-660,* which provides a measure of information processing and spatial orientation (PILAPT Handbook, 2006).

Information Processing. This group included five tests. (a) *Attention-670* is an information processing test requiring rapid apprehension of the up-down and left right dimension. (b) *Determinastionsgerät (DTG-690)* measures reactive stress tolerance, attention deficits, and reaction speed in the presence of rapidly changing and continuous optical and acoustic stimuli. (c) *Digit Recall (DT-680)* is a test of short-term memory, in which the candidate is presented with a set of numbers of varying size. Once the numbers disappear, the candidate is required to type them in. (d) *Numbers-700* measures perceptual speed and is associated with quick, accurate, and efficient use of visual perception, as well as working memory. (e) *Vigilance-720* measures attention, and the candidate has to complete two distinct tasks, a routine task and a priority task.

Personality Traits and Interview Ratings

Personality Traits. Personality Traits were assessed with the 5PFmil 2.0 (Five Personality Factors, Military version 2.0) (Engvik, 1993; 1997). The measure included a total of 240 statements related to the Big-five Traits (Extroversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to New Experiences). The factor structure and psychometric properties have been supported in previous studies (Friborg, Barlaug, Martinussen, Rosenvinge, & Hjemdal, 2005). Participants rated the items (e.g., "I am someone who is outgoing" or "I am someone who can be tense") on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*). The mean score for each Big-five Trait was converted to a T-score based on norms from the Norwegian Armed Forces. Cronbach's alpha for the Big-five traits were for Extroversion (.80), Agreeableness (.83), Conscientiousness (.84), Emotional Stability (.79), and Openness to New Experiences (.66) based on this sample.

Interview Ratings. The interview was semi-structured and conducted by a psychologist and a pilot together (one hour). The aim is to assess motivation, previous experience and competencies and addition to personality characteristics. The psychologist has access to the test results of each candidate including results from a Big-five measure when conducting the interview. Interview ratings were assigned based on interview performance and test results, which were summarized in two scores (Stanine): School Prognosis and Pilot Prognosis. A Total Prognosis was also calculated as the sum of School and Pilot Prognosis.

Pilot Training Performance Criteria

Several criteria collected during the three-year bachelor program were included in this evaluation.

Pre-Flight Ground School Theory Mean Score. A total of nine multiple-choice exams were included in the Pre-Flight Ground School Theory (PFGS) Mean Score, with each exam scored in terms of percent correct answers. The exams were based on basic aviation theory from the second year with nine topics, including aviation law, aircraft general knowledge, flight performance and planning, human performance, meteorology, navigation, operational procedures, principles of flight, and VFR communication.

Air Traffic Pilot License Theory Mean Score. The Air Traffic Pilot License Theory (ATPL) Mean Score included 14 theoretical exams (e.g., air law, meteorology, general navigation, principles of flight, aircraft general knowledge, systems, performance). The score was calculated as the mean percent correct answers, similar to the Pre-Flight Ground School Theory Mean Score.

Civil Aviation Authorities Air Traffic Pilot License Theory Mean Score. After completing the ATPL exam at the UiT, students may take the corresponding exam with the Norwegian Civil Aviation Authorities. It includes a total of 14 exams, and candidates must receive a score of at least 75% on each exam to pass. The score used in the analyses is the mean percent correct answers based on a total of 14 exams.

Extra Flying Hours Needed Above Minimum Requirements. The training manual stipulates the minimum number of flying hours required to complete the bachelor program. Since the program has undergone some changes over the years, including the minimum number of flying hours needed, Extra Flying Hours Needed was calculated as the percent of hours over the minimum the participant had to complete relative to the minimum number of flying hours at the time. A high score indicated poor performance as extra flying hours were needed. Practical flight training was conducted at a different flight school at the beginning of the study period, and these results were not available for this evaluation, resulting in a lower sample size for analyses involving this criterion

Mean University Grade (academic exams). Mean University Grade was based on six subjects (philosophy of science, mathematics, physics, organization and leadership, business economy, and the bachelor thesis). Courses were graded from A (best score) to F (failure); these grades were converted to numbers with values from 0 to 5, where 5 corresponds to a grade of A.

Statistical Analyses

The statistical analyses were performed with the Statistical Package for Social Sciences (SPSS 26) including calculation of descriptive statistics and hierarchical regression analyses. Correlations were corrected for multivariate range restriction (Lawley, 1944) using the Rangej software (Johnson & Ree, 1994). Correlations were categorized as small (.10), medium (.30), and large (.50), in line with Cohen’s suggestions (Cohen, 1988). Hierarchical regression analyses were computed to predict three performance criteria: Aviation Theory (overall mean of the scores for PFGS, ATPL, CAA ATPL), Extra Flying Hours Needed and Mean University Grade. In these analyses, the PFGS, ATPL, CAA ATPL Mean Scores were combined to an overall mean score (Aviation Theory) used as criterion, in addition to predicting the criteria Extra Flying Hours Needed, and Mean University Grade. The predictors were entered in three steps in the same order as in the selection process with GMA in the first step (paper-and-pencil tests), the computer-based cognitive and psychomotor tests in the second step, and interview ratings in the third step. Big-Five Traits were not added to the model, as there is no explicit selection on these measures; instead this information is integrated into the interview ratings using clinical judgement.

Results

The Predictive Validity of Paper-and-Pencil and Computer-Based Tests

Bivariate correlations (uncorrected and corrected) revealed that GMA tests were significantly related to two of the aviation theory-related performance criteria (PFGS and ATPL Mean Score), but not to Mean University Grade as expected. The three groups of computer-based tests were all significantly related to most aviation-related performance criteria, including Extra Flying Hours Needed, which is the most skill-based criterion. The Total Test Score (based on all 16 tests) was a significant predictor of all performance criteria except Mean University Grade. The significant uncorrected correlation was -.22 (r_c = -.33) between Total Test Score and Extra Flying Hours Needed, and ranged from .27 to .38 between Total Test Score and PFGS, ATPL, and CAA-ATPL Mean Score (r_c = .40 – .54) (Table 1).

Insert Table 1

The Predictive Validity of Personality Traits and Interview Ratings

The majority of the bivariate correlations between the Big-five Traits and the five performance criteria were small and non-significant (Table 2). However, Conscientiousness was related to three of the criteria, with uncorrected correlations between .17 and .23 (r_c = .18 and .25), and Emotional Stability was significantly related to PFGS .18 (r_c = .21). The interview ratings were significantly related to all performance criteria, except for Extra Flying Hours Needed, which showed medium to large correlations (Table 2).

Insert Table 2&3

A Model for Predicting Pilot Training Performance Criteria

Overall, the hierarchical regression model explained 25% of the variance in Aviation Theory, 19% of Extra Flying Hours Needed, and 7% of Mean University Grade (Table 3). In general, the first step (paper-and-pencil GMA tests) explained a small, and for two of the criteria a non-significant, part of the variance. The second step (computer-based cognitive and psychomotor tests) explained a significant part of the variance in Aviation Theory and Extra Flying Hours Needed, and the third step (interview ratings) added incremental validity to the prediction of Aviation Theory and Mean University Grade, but not Extra Flying Hours Needed.

Discussion

This study examined how well the test battery and selection system at the UiT predicted a variety of performance criteria collected during the pilot training and academic studies. GMA tests were significantly related to two of the aviation theory-related performance criteria, but not to Mean University Grade. In general, the predictive validity of the paper-and-pencil GMA tests was somewhat lower than that of the computer-based tests, which is in line with previous meta-analyses (Hunter & Burke, 1994; Martinussen, 1996), in which measures of GMA had lower mean predictive validities than more specific cognitive and psychomotor tests. Overall, the observed correlations were higher than those reported previous studies conducted by the Norwegian Air Force, which uses the same test battery (Martinussen & Torjussen, 2004; Lang-Ree, Martinussen, & Ødegaard, 2010). This may be explained by possible differences in the selection ratio or the use of more reliable performance criteria in our study. Moreover, our criteria were collected over a much longer time period and included official exams, which probably have a better reliability than instructor ratings. The military studies mostly used pass/fail as a measure of success in initial training, which may be contaminated by other factors such as lack of motivation or illness, and in most cases, they assessed performance over a relatively short time period. Our findings are also in line with previous job analyses indicating that many cognitive and psychomotor abilities were described as relevant or highly relevant by experienced pilots (Goeters, Maschke, & Eißfeldt, 2004).

Ever since the beginning of pilot selection, personal qualities, as well as cognitive and psychomotor skills, have been regarded as important in order to become a good pilot (Dockeray & Isaacs, 1921; Hunter, 1989). Cooperative and social skills have also been highlighted in job analyses of civilian pilots and in recent recommendations from EASA (2018a,b). In our selection system, non-technical skills and personality traits and characteristics were assessed during the interview, during which the psychologist had access to the Big-five profile. The overall interview rating predicted all the performance criteria except Extra Flying Hours Needed, which is the criterion which most directly assesses flying performance. There were small differences between the two interview ratings (School Prognosis and Pilot Prognosis) in terms of predictive validity, and overall, Total Prognosis had better predictive validity than School or Pilot Prognosis for all performance criteria.

The regression models indicated that the interview ratings added incremental validity beyond the cognitive tests for two of the performance criteria. This is in line with a recent unpublished meta-analysis which indicated that the interview may add incremental validity beyond the GMA tests (Schmidt, Oh, & Schaffer, 2016). The Big-five Traits had small, non-significant correlations with the performance criteria, except for Conscientiousness and Emotional Stability, which were related to two and one of the aviation theory-related criteria, respectively. This corresponds to previous findings from meta-analyses and single studies in which the overall predictive validity of personality traits was small for pilot selection

(Campbell et al., 2009; Carretta et al., 2014, Martinussen, 1996). Our findings also correspond to meta-analytic results indicating that the best overall predictors of work performance are Conscientiousness and Emotional Stability, but it varies with the type of criteria and occupation examined (Salgado, 1998; Salgado, & Táuriz, 2014). A study of military aviator performance indicated that Agreeableness and Extroversion were correlated with more favorable supervisor-rated performance for both pilots and navigators (Barron, Carretta, & Bonto-Kane, 2016), and the authors argued that the reason for the low predictive validities in pilot selection was the choice of criterion, which was mostly based on training performance rather than actual work performance. The criteria used in our study were mostly related to individual performance, not team-related or actual pilot performance in a work setting, which may explain the low validity coefficients for personality traits in our study. However, the use of the Big-five Traits for selection purposes has been criticized (see e.g., Morgeson, Campion, Dipboye, Hollenbeck, & Schmitt, 2007), and at the moment, there is little evidence supporting the explicit use of cut-off scores for the Big-five scales. Information about personality traits may be useful to the psychologist during the interview, and our results showed that interview ratings demonstrated low to medium correlations with Big-five Traits. The highest correlations were found between Extroversion and the Total Prognosis (r = .24) which indicated that extroverted candidates received a higher interview rating. Another issue is of course related to the possibility of biases in self-evaluations, such as responding in a socially desirable way or simply lacking insight into one's own traits.

Study Limitations

One limitation of this study is that it was conducted in a university setting, where the majority of criteria were exam results and not operational performance in an airline. However, it can be very difficult to collect information on criteria after graduation, in addition to the fact that the work situation and type of job may not be comparable between pilots. For the university, it is of course relevant to predict student performance, but a more standardized measure of direct flying performance would have improved the study. There were some missing data in our sample, mostly due to a lack of data from the first students who entered the program, which reduced the statistical power of some of the analyses.

Conclusion and Practical Implications

An important aspect of the present study was to examine the predictive validity of performance training criteria and to examine the selection system as part of an evidence-based practice in personnel selection. In general, the findings confirmed the predictive validity of both the paper-and-pencil GMA tests and the computer-based cognitive and psychomotor selection tests. The interview had incremental validity for most of the investigated performance criteria, but it may benefit from being more structured as indicated in meta-analyses and reviews of the employment interview (Levashina, Hartwell, Morgeson, & Campion, 2014; Schmidt et al., 2016). Increasing standardization in terms of both content, how the responses are evaluated, as well as training interviewers may increase both reliability and the predictive validity of the interview (Levasia et al., 2014). Another approach to improving the selection system would be to improve the reliability of both the selection tests and the performance criteria, but also to identify domains that are not assessed in the current selection system. Job analyses and predictions of future tasks for commercial pilots have indicated an increase in requirement levels, including an even stronger emphasis on visualization (Bruder et al., 2013; Eißfeldt, et al., 2009). Future improvements could include supplementing the existing test battery with tests of currently unassessed cognitive abilities, such as visualization, monitoring and decision making. Also, developing more long-term criteria on actual job performance after graduation, may provide a better evaluation of the personality traits as indicated in a study of military aviator performance (Barron et al., 2016).

References

ALMamari, K., & Traynor, A. (2019). Multiple test batteries as predictors for pilot performance: A meta-analytic investigation. *International Journal of Selection and Assessment, 27*(4), 337-356. https://doi.org/10.1111/ijsa.12258

ALMamari, K., & Traynor, A. (2020). Predictive validity of the Air Force Officer Qualification Test (AFOQT) for pilot performance. *Aviation Psychology and Applied Human Factors*, *10*(2), 70-81. https://doi.org/10.1027/2192-0923/a000190.

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Barron, L. G., Carretta, T. R., & Bonto-Kane, M. V. A. (2016). Relations of personality traits to military aviator performance: It depends on the criterion. *Aviation Psychology and Applied Human Factors, 6*(2), 57–67. https://doi.org/10.1027/2192-0923/a000100

Broach, D., Schroeder, D., & Gildea, K. (2019). *Best practices in pilot selection.* Report no. DOT/FAA/AM-19/6. Office of Aerospace Medicine, Washington, DC. https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/2010s/media/201906.pdf

Bruder, C., Eißfeldt, H., Grasshoff, D., Friedrich, M., Hasse, C., Hoermann, H.-J., Hoff, A., Papenfuß, A., Schulze Kissing, D., Uebbing-Rumke, M., Wenzel, J., & Zierke, O. (2013). *Simulator-based research on operational monitoring and decision making for human operators in future aviation.* Project Report Aviator II. DLR, Hamburg.

Burke, E. F., Kitching, A., & Valsler, C. (1997). *The Pilot Aptitude Tester (PILAPT): On the development and validation of a new computer-based test battery for selecting pilots.*

Proceedings of the 9th International Symposium on Aviation Psychology. Columbus, OH: State University.

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. https://doi.org/10.1080/10508410903415872

Carretta, T. R. (2011). Pilot Candidate Selection Method. Still an effective candidate predictor for US Air Force pilot training performance. *Aviation Psychology and Applied Human Factors, 1*(1), 3-8. https://doi.org/10.1027/2192-0923/a00002

Carretta, T. R., & King, R. E. (2020). History of pilot selection. In R. Bor, C. Eriksen, Hubbard, T. P., & R. King, *Pilot selection. Psychological Principles and Practice* (pp. 9-20). CRC Press.

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). Lawrence Erlbaum Associates.

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 perforamance. *The International Journal of Aviation Psychology, 24*(4), 247-264. http://dx.doi.org/10.1080/10508414.2014.949200

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

Damos, D. (2011). *KSAOs for military pilot selection: A review of the literature*. Report number AFCAPS-FR-2011-0003. Randolf AFB, TX: Air Force Personnel Center Strategic Research and Assessment.

Dockeray, F. C., & Isaacs, S. 1921. Psychological research in aviation in Italy, France, England, and the American Expeditionary Forces. *Journal of Comparative Psychology, 1*, 115–148.

European Union Aviation Safety Agency (EASA). (2018a, July 23). COMMISSION REGULATION (EU) 2018/1042 (CAT.GEN.MPA.175). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A32018R1042&from=EN

European Union Aviation Safety Agency (EASA). (2018b, November 28). Annex III to Decision 2018/012/R 'AMC and GM to Part-CAT — Issue 2, Amendment 15'. Retrieved from https://www.easa.europa.eu/sites/default/files/dfu/Annex%20III%20to%20EDD%202018-012-R.pdf

Eißfeldt, H. (2014). Commentary on the article by King: Select in/select out- what aviation psychology offers for pilot selection. *The International Journal of Aviation Psychology, 24*(1), 78-81. https://doi.org/10.1080/10508414.2014.860840

Eißfeldt, H., Grasshoff, D., Hasse, C., Hoermann, H-J., Kissing, D. S., Stern, C., Wenzel, J., & Zierke, O. (2009). *Aviator 2030. Ability Requirements in Future ATM Systems II: Simulations and Experiments*. Hamburg: Deutsches Zentrum für Luft- und Raumfahrt e. V. Institut für Luft- und Raumfahrtmedizin, Luft- und Raumfahrtpsychologie. Retrieved from http://www.dlr.de/me/Portaldata/25/Resources/dokumente/Aviator_2030_Report_FB_2009-28.pdf

Engvik, H. (1993). «Big Five» på norsk [Big five in Norwegian]. *Journal of the Norwegian Psychologists Association, 30,* 884-896.

Engvik, H. (1997). *5PFmil 2.0* (Computer software). Department of Psychology, University of Oslo, Norway.

European Federation of Psychologists' Association (EFPA). (2013). *EFPA Review model for* *the description and evaluation of psychological tests: Test review form and notes for reviewers, v 4.2.6*. Retrieved from http://www.efpa.eu/professional-development

Friborg, O, Barlaug. D, Martinussen, M., Rosenvinge, J. H., & Hjemdal, O. (2005). Resilience in relation to personality and intelligence. *International Journal of Methods in Psychiatric Research*, *14*(1), 29-42. https://doi.org/10.1002/mpr.15

Goeters, K. M., Maschke, P., & Eißfeldt, H. (2004). Ability requirements in core aviation professions: Job analysis of airline pilots and air traffic controllers. In K. M. Goeters (Ed.), *Aviation psychology: Practice and Research* (pp. 99–119). Ashgate Publishing Limited.

Hörmann, H. J., & Goerke, P. (2014). Assessment of social competence for pilot selection. *The International Journal for Aviation Psychology, 24*(1), 6-28. https://doi.org/10.1080/10508414.2014.860843

Hörmann, H. J., & Maschke, P. (1993). *Personality scales as predictors of job success of airline pilots*. In R. S. Jensen (Ed.), Proceedings of the Seventh International Symposium on Aviation Psychology (pp. 450-454). Columbus: The Ohio State University.

Hörmann, H. J., & Maschke, P. (1996). On the relation between personality and job performance of airline pilots. *International Journal of Aviation Psychology, 6,* 171–178. https://doi.org/10.1207/s15327108ijap0602_4

Hunter, D. R. (1989). Aviator selection. In *Military personnel measurement: Testing, assignment, evaluation,* M. F. Wiskoff & G. F. Rampton (Eds.) (pp. 129–167). Praeger.

Hunter, D. R., & Burke, E. F. (1994). Predicting aircraft pilot-training success: A meta-analysis of published research. *International Journal of Aviation Psychology, 4*(4)*,* 297-313. https://doi.org/10.1207/s15327108ijap0404_1

International Civil Aviation Organization (IATA). (2019). *Pilot aptitude testing. Guidance* *material and best practices for pilot aptitude testing* (3rd Ed.). Montreal, Canada. https://www.iata.org/contentassets/19f9168ecf584fc7b4af8d6d1e35c769/pilot-aptitude-testing-guide.pdf

Johnson, J. F., Barron, L. G., Carretta, T. R., & Rose, M. R. (2017). Predictive validity of spatial ability and perceptual speed test for aviator training. *The International Journal of Aerospace Psychology, 27*(3-4), 109-120. https://doi.org/10.1080/24721840.2018.1442222

Johnson, J. T., & Ree, M. J. (1994). Rangej: A Pascal program to compute the multivariate correction for range restriction. *Educational and Psychological Measurement, 54*(3), 693-695.

Lang-Ree, O. C. (2021). *Test-retest reliability estimates of pilot selection tests: A small scale study*. Internal short report. Norwegian Armed Forces Joint Medical Services, Oslo, Norway.

Lang-Ree, O. C., Martinussen, M., & Ødegaard, P. E. (2010, September 20-24). *Pilot selection in the Norwegian Air Force: A validation study* [Poster presentation]. European Association for Aviation Psychology 29th Conference, Budapest, Hungary.

Lawley, D. (1944). IV.—A Note on Karl Pearson's Selection Formulae. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences, 62*(1), 28-30. https://doi.org/10.1017/S0080454100006385

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014), The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology, 67*(241-293). https://doi.org/10.1111/peps.12052

Martinussen, M. (1996). Psychological measures as predictors of pilot performance: A meta-analysis. *International Journal of Aviation Psychology, 6*(1), 1-20. https://doi.org/10.1207/s15327108ijap0601_1

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. https://doi.org/10.1207/s15327108ijap0801_2

Martinussen, M. (2016). Pilot selection. An overview of aptitude and ability assessment. In R. Bor, Eriksen, C., Oaks, M., & P. Scragg, *Pilot mental health assessment and support* (pp. 23-39). Routledge Taylor and Francis Group.

Martinussen, M., & Torjussen, T. M. (2004). Initial validation of a computer-based assessment battery for pilot selection in the Norwegian Air Force. *Human Factors and Aerospace Safety, 4*(3), 233-244.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x

Paullin, C., Katz, L., Bruskiewicz, K. T., Houston, J., & Damos, D. (2006). *Review of aviator selection*. Technical report 1183. United States Army Research Institute for the Behavioral and Social Sciences.

Raven, J. C. (1986). *Advanced Progressive Matrices*. Oxford: Psychologist Press.

Salgado, J. F. (1998). Big Five personality dimensions and job performance in army and civilian occupations: A European perspective. *Human Performance*, *11*(2), 271–288. doi:10.1207/s15327043hup1102&3_8

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3-30. doi:10.1080/1359432X.2012.716198

Schmidt, F. L., Oh, I.-S., & Shaffer, J. A. (2016). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research findings.* Working paper. doi: 10.13140/RG2.2.18843.26400

The Pilot Aptitude Tester (PILAPT) and Controller Aptitude Tester (CONAPT) Manual. (2006). London: Peoples' Technologies Ltd.

Zierke, O. (2014). Predictive validity of knowledge tests. *Aviation Psychology and Applied Human Factors, 4*(1), 98-105. https://doi.org/10.1027/2192-0923/a000061

Table 1. *Bivariate Correlations between Groups of Tests and Pilot Training Performance Criteria* (*N* = 160-180)

Variables	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
Groups of Tests													
1. GMA	6.26	0.99	-						*.35*	*.36*	*.32*	*-.00*	*.13*
2. Psychomotor Coordination	6.23	1.46	-.05	-					*.39*	*.31*	*.31*	*-.36*	*.03*
3. Spatial Ability	6.28	1.29	.18*	.37**	-				*.49*	*.41*	*.36*	*-.42*	*.04*
4. Information Processing	6.00	1.03	.38**	.16*	.36**	-			*.45*	*.39*	*.28*	*-.31*	*.11*
5. CBT Total (tests 2-4)	6.14	0.86	.25**	.68**	.85**	.75**	-		*.54*	*.46*	*.39*	*-.44*	*.09*
6. Total Test Score (tests 1-4)	6.21	0.76	.62**	.53**	.69**	.76**	.96**	-	*.54*	*.48*	*.40*	*-.33*	*.12*
Criteria													
7. PFGS Mean Score	86.05	6.28	.18*	.23**	.32**	.30**	.38**	.38**	-				
8. ATPL Mean Score	88.49	4.55	.22**	.15	.25**	.26**	.30**	.33**	.74**	-			
9. CAA ATPL Mean Score[a]	92.25	3.41	.18	.18	.21*	.15	.25**	.27**	.46**	.67**	-		
10. Extra Flying Hours Needed (above min.)[a,b]	10.28	7.33	.15	-.28**	-.34**	-.17	-.35**	-.22*	-.30**	-.46**	-.35**	-	
11. Mean University Grade[c]	3.39	0.86	.11	-.01	-.01	.09	.04	.08	.52**	.58**	.46**	-.34**	-

Note. [a] *N* = 106-110. [b]Calculated as percent additional flying hours above minimum requirements to complete the program. [c]Mean grades based on theoretical university exams. GMA = General Mental Ability. CBT = Computer-based tests. PFGS = Pre-Flight Ground School Theory. ATPL = Air Traffic Pilot License Theory. CAA ATPL = Civil Aviation Authorities Air Traffic Pilot License Theory. The correlations (in italics) above the diagonal are corrected for multivariate range restriction. *$p < .05$. **$p < .01$ (two-tailed).

Table 2. *Bivariate Correlations between Big-five Traits, Interview Ratings, and Pilot Training Performance Criteria* (*N* = 153-175)

Variables	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
Big-five Traits															
1. Extroversion	53.73	7.68	-								*.05*	*.06*	-*.02*	*.07*	*.02*
2. Agreeableness	48.19	9.46	.57**	-							*.07*	*.12*	-*.01*	-*.01*	*.06*
3. Conscientiousness	56.22	6.87	.47**	.47**	-						*.22*	*.25*	-*.06*	*.00*	*.18*
4. Emotional Stab.	51.74	6.87	.51**	.43**	.41**	-					*.21*	*.18*	-*.07*	-*.09*	*.04*
5. Openness	48.41	8.31	36*	.51**	.14	.27**	-				*.09*	*.10*	*.00*	*.10*	*.08*
Interview Ratings								-							
6. School Prognosis	6.92	1.11	.13	.05	.10	.13	.17*		-		*.50*	*.53*	*.64*	-*.14*	*.34*
7. Pilot Prognosis	6.71	1.32	.27**	.19**	.11	.19**	.10	.41**			*.53*	*.53*	*.51*	-*.28*	*.36*
8 Total Prognosis	13.63	2.04	.24**	.15*	.12	.19**	.16*	.81**	.87**	-	*.53*	*.55*	*.63*	-*.22*	*.35*
Criteria															
9. PFGS Mean Score	86.05	6.28	.03	.05	.20*	.18*	.09	.33**	.34**	.39**	-				
10. ATPL Mean Score	88.49	4.56	.04	.10	.23**	.15	.10	.36**	.34**	.41**	.74**	-			
11. CAA ATPL Mean Score[a]	92.25	3.41	-.01	.00	-.05	-.06	.00	.50**	.31**	.48**	.46**	.67**	-		
12. Extra Flying Hours Needed [a,b]	10.28	7.34	.06	.00	.01	-.07	.09	-.05	-.18	-.15	-.30**	-.46**	-.35	-	
13. Mean University Grade [c]	3.39	0.86	.01	.06	.17*	.03	.08	.21**	.22**	.25**	.52**	.58**	.46**	-.34**	

Note. [a]*N* = 101 -110. [b]Calculated as percent additional flying hours above minimum requirements to complete the program. [c]Mean grades based on theoretical university exams. PFGS = Pre-Flight Ground School Theory. ATPL = Air Traffic Pilot License Theory. CAA ATPL = Civil Aviation Authorities Air Traffic Pilot License Theory. The correlations (in italics) are corrected for multivariate range restriction. **p* < .05. ***p* < .01 (two-tailed).

Table 3. *Hierarchical Multiple Regression Analysis for Predicting Pilot Training Performance Criteria*

Variables	Aviation Theory (mean)[a]		Mean University Grade[c]		Extra Flying Hours Needed[b]	
	β	Δ R^2	*β*	Δ R^2	*β*	Δ R^2
Step 1. Paper-and-pencil tests						
General Mental Ability	.07	.04**	.07	.01	.26**	.02
Step 2. Computer-based tests						
Mean score (Psychomotor, Spatial and Info.processing)	.21**	.10**	-.07	.00	-.39**	.16**
Step 3. Interview ratings						
Mean score (School and Pilot Prognosis)	.36**	.11**	.26**	.06**	-.07	.00
Total R^2		.25**		.07**		.19**
N	169		179		104	

Note. [a]Mean results from PFGS, ATPL, and CAA ATPL. [b]Calculated as percent additional flying hours above minimum requirements to complete the program. [c] Mean grades based on theoretical university exams. All coefficients were taken from the last step of the equation.

*p < .05. **p < .01.