A Meta-analysis on Air Traffic Controllers Selection: Cognitive and Non-Cognitive Predictors

Damien Mouratille[1,2], Franck Amadieu[2], & Nadine Matton[1,2]

[1] ENAC, 7 avenue Edouard Belin, 31055, Toulouse, France

[2] Laboratoire CLLE, CNRS 5263, University Toulouse II Jean Jaures, Toulouse, France

Author Note

Correspondence concerning this article should be addressed to Damien Mouratille, ENAC, 7 avenue Edouard Belin, 31055, Toulouse, France. E-mail: damien.mouratille@enac.fr

Abstract

This psychometric meta-analysis investigated the relation of cognitive and non-cognitive factors to the training success of Air Traffic Controllers by synthesizing 51 studies (N = 65839). Cognitive factors were classified by Cattel-Horn-Carrol theory. Cognitive composite scores and work samples were also included. Non-cognitive factors consisted of Big Five personality traits, biodata, motivation and non-cognitive composite scores. Medium effect was measured for cognitive factors ($k = 45$, $p = .37$). Quantitative knowledge, processing speed, work sample, short-term working memory, cognitive composite and visuo-spatial processing predictors showed large effects ($p > .30$). Significant moderating effects of criterion nature and period of publication were observed. Initial training (k = 30, $p = .50$) was generally better predicted than on-the-job training (k = 25, $p = .18$). Better predictive validity was measured from the 60's to nowadays. For non-cognitive factors, only a small effect was measured (k = 24, $p = .15$). Non-cognitive composites and education showed large effects ($p > .30$). No significant relation was measured between Big Five personality traits and success criteria. The present findings suggest that selection processes used for Air Traffic Controllers should focus on cognitive predictors or other methods of assessments. Data and scripts can be found at https://osf.io/mkyw7/.

Keywords: Meta-Analysis, Air traffic controllers, training success, cognitive predictors, non-cognitive predictors

Word count: 8671

A Meta-analysis on Air Traffic Controllers Selection: Cognitive and Non-Cognitive Predictors

Introduction

Air traffic is expected to increase significantly during the next twenty years. With the traffic recovery post-covid-19, a growth of at least 1.9 percent per year is foreseen for the period 2021 to 2041 (FAA, 2021). To meet this increase, air traffic controller recruitment is clearly a public priority since a significant shortage have been forecast (International Civil Aviation Organization, 2011).

Air traffic controllers (ATCOs) are responsible for the organization of air traffic to ensure its safety and efficiency. Since ATCOs manage all aircraft in their sectors, the growth of traffic will increase sector complexity and ATCO workload. An excessive workload can be detrimental to safety and before a limit is reached, one solution would be to recruit new ATCOs. In Eurocontrol member states, ATCO training takes approximately 3 years and costs several hundreds of thousands per student depending on the Air Navigation Service Provider (ANSP). For example, all ATCOs in France and in the member nations of European Civil Aviation Conference (ECAC) have to validate initial training and then start unit training before being fully operational. Initial training is based on seven phases incorporating theory and practical training on high-fidelity ATC simulators whereas unit training is mainly practical with real sector on simulator and on-the-job training. The entire process from initial training to on-the-job training takes around five years plus or minus 2. Given the time and cost of training before students become fully certified ATCOs, special attention is devoted to selection and training processes. Attrition (failure or drop-out) is a burden for both institutions and students: psychological and possibly financial burden for the student and loss of time and money for the institution. Attrition can be explained by numerous reasons but the main one is that training is challenging and not everyone has the necessary abilities. The profession is cognitively demanding and stressful and, consequently, training difficulty is related to job difficulty.

Day-to-day activities are mainly related to the routing of aircraft while adapting to changing conditions. Operational requirements can change rapidly and unpredictably, escalating issues arise very quickly with and without the controller's actions: the weather can change from the absence of wind to a thunderstorm, an aircraft can have a technical failure, a military zone can be activated, a pilot may commit an error and not follow the ATCO's clearances, an entire national airspace can be closed such as after September 11 attacks or United Kingdom ANSP system failure in 2014, etc.

In order to maintain separation between aircraft while receiving the requests of pilots and handling them efficiently, ATCOs must create a mental image of the environment by perceiving the situation, anticipating the future, and taking decisions. This mental image is built from multiples sources of information: long-term memory such as regulations and maps, radio communications, radar displays, decision support tools, colleagues, and/or the view outside the tower. Their level of responsibility is very high as their errors may fatally contribute to aviation accidents: for instance, one of the main contributing factors of the 2002 Uberlingen mid-air collision was the late conflict detection by the Skyguide ATCO (Brooker, 2008). During training, students are prepared to manage all this unpredictability.

The terms of selection process are determined mainly by the ANSP. The purpose of the process is to select candidates with a strong probability of training success. A wide variety of methods can be used (Martinussen & Hunter, 2017): knowledge tests, biographic data, personality assessment, cognitive testing, work-sample, situational judgement tests, and/or job interview. Each method has its advantages and shortcomings but almost every ANSP uses plural methods. The ANSP selection process can be decomposed into one or more steps and psychotechnical testing is mainly used in order to assess individual differences in cognitive abilities and personality. For example, the Federal Aviation Administration (FAA), one of the major actors in selection psychology (Bleckley, Crutchfield, King, Manning, &

Carretta, 2009), used to employ an eight-hour computer-based battery (Air Traffic Selection And Training). In 2016, FAA replaced this eight-hour battery with a three-hour computer-based battery (Air Traffic Skill Assessment) that comprises seven cognitive tests. With 14375 ATCOs on duty during the 2019 fiscal year and recruitment of around 1000 ATCOs per year (Federal Aviation Administration, 2020), attrition is such a major concern for FAA that the selection process is frequently audited. In 2020, a new audit was mandated by the US Congress on the ATCO pre-employment test validity (U.S. Department of Transportation, 2020). In Europe, the two countries with the highest number of ATCOs (Germany and France) use different processes. German ATCOs are selected with tests developed by the German Aerospace Center DLR (Pecena, Gayraud, & Eißfeldt, 2019). They assess cognitive abilities, English knowledge, teamwork abilities, and motivation. On the contrary, French ATCOs are only selected on non-aeronautical knowledge tests (mathematics, French, English, physics) but must have completed two years of Higher School Preparatory Classes or university. Since 2003, Eurocontrol has offered a computer-based test battery known as the First European Air traffic controller Selection Test (FEAST). Initially available to all ECAC member states, FEAST is now used by 50 organizations worldwide (mainly in Europe) and 110.000 candidates have been tested on criteria such as cognitive abilities, knowledge, English, work-sample and personality (Eurocontrol, 2020). Germany, France, and countries using FEAST have different selection processes but finally, European ATCOs all have the same European ATCO student License which allows a controller to work for a foreign ANSP. These four selection processes are a small sampling of all selection processes in the world but are a representative sample.

In personnel selection, an efficient selection process must have an empirical basis to identify variables that are valid and reliable predictors of student success (Lievens, Sackett, & Zhang, 2020; Roe, 2005). However, when we look at the predictive validity of these selection processes, much variability can be seen and yet little has been published on the subject. Some authors have published scientific reviews on the ATCO selection process (Broach, 2017;

Broach & Manning, 1997; King et al., 2007; Pavel, 2012) but a limitation shared by these studies is the generalization of the results. A predictor can be valid with one entity during a certain period, but the real question is rather to know if the predictor is valid all the time and with all entities with their own particularities. The latest meta-analysis on ATCO selection was carried in 2000 (Martinussen, Jenssen, & Joner, 2000). From 35 independent samples, the meta-analysis highlighted low-to-medium mean correlations between .03 and .27: .03 for personality predictor, .18 for occupational knowledge, .19 for mathematical ability and multiple task performance, .21 for verbal and general abilities, .25 for spatial ability, .28 for work sample and .30 for composite scores. All were significantly positive across all studies except for the personality predictor. Psychotechnical assessments can be separated into two parts: one part assesses the cognitive aspect and the other the non-cognitive aspect.

Cognitive predictors

The impact of cognitive abilities on occupational performance has been widely investigated. Today, Cattel-Horn-Carrol (CHC) theory (Schneider & McGrew, 2018) is the dominant theory for classifying specific cognitive abilities (McGill & Dombrowski, 2019; for a contrary view, Wasserman, 2019). CHC theory is born from the merging of a two-dimensional theory (Cattell, 1963; Horn, 1965) and a three-dimensional one (Carroll, 1993; for more details on the history of intelligence theory, see Flanagan & McDonough, 2018). CHC model breaks down intelligence into a three-level hierarchical structure: the highest level (Stratum III) is composed of general cognitive ability (*g* factor), the next level (Stratum II) of 18 broad cognitive abilities, and the lowest level (Stratum I) of more than 80 narrow cognitive abilities. The broad cognitive abilities of interest here are quantitative knowledge, reading and writing, comprehension-knowledge, domain-specific knowledge, processing speed, fluid reasoning, short-term memory, long-term storage and retrieval, visuo-spatial processing, and ability emotional intelligence. The *g* factor is known to be the best predictor of job performance: Schmidt and Hunter (1998) showed general mental ability had a high validity in general (r = .51). Lang and Kell (2020) revealed that the g factor could predict professional success even 51 years after taking the tests and was the dominant predictor in incremental validity analyses. In relative importance analyses, specific cognitive abilities sometimes explained more variance than the g factor. A "great debate" is being held on general versus specific abilities (Beier, Kell, & Lang, 2019). Recently, ALMamari and Traynor (2019) employed CHC theory for classifying special cognitive predictors (composite score, see below) for pilot performance.

Some tests are designed to mainly assess one broad cognitive ability but in other tests, known as work sample, many broad cognitive abilities are assessed since the focus of assessment is the ecological aspect or higher cognitive functions. Work sample tests are closer to the real job performed and some of them can assess non-cognitive abilities. In ATCO selection, work samples are mainly designed to assess the cognitive component of the job. In a meta-analysis, Roth, Bobko, and McFarland (2005) estimated that work sample tests had a mean validity of .33. In addition, composite scores are used in selection psychology: multiple test scores combined and weighted into one score (Hattrup, 2012; for example, see ALMamari & Traynor, 2019). One limit of the composite score is its validity: it can be a type of black box or "rotten pot" (Bobko, Roth, & Buster, 2007). As highlighted by ALMamari and Traynor (2019), "it is a combination of skills and competencies that jointly contribute to the scores given to applicants and influence organization's employment decisions". Nevertheless, this combination sometimes is not meaningful.

Non-cognitive predictors

For the non-cognitive aspect (Gutman & Schoon, 2013), four constructs are generally considered: personality, emotional intelligence, motivation, and biodata.

Personality refers to the characteristic sets of behavior, cognition, and emotional patterns that evolve from biological and environmental factors (Corr & Matthews, 2020). The dominant model used for personality assessment is the Big Five (Costa & McCrae, 1999). Five factors have been proposed: Openness to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism. In the military context, conscientiousness was the best personality predictor for job performance, extraversion for training performance and neuroticism for counter-productive work behavior (Darr, 2011). Emotional intelligence is regarded “as a distinct construct from traditional IQ and personality, which facilitates the potential for prediction of and influence on various real-life outcomes” (Fiori & Vesely-Maillefer, 2018). Defined for the first time by Salovey and Mayer (1990), emotional intelligence is “the ability to monitor one’s own and others’ feelings, to discriminate among them, and to use this information to guide one’s thinking and action”. During the following decade, emotional intelligence has been divided into two research streams: ability and trait. Ability emotional intelligence is composed of four emotion-related abilities: perceiving, using, understanding, and managing emotion (Mayer & Salovey, 1997). Trait emotional intelligence is “a constellation of emotional perceptions located at the lower levels of personality hierarchies” (Petrides & Furnham, 2000; Petrides, Pita, & Kokkinaki, 2007) and is assessed through self-reports. For ability emotional intelligence, problems have to be solved and accuracy of solution is the focus whereas for trait emotional intelligence, the belief of adapting to emotional situations is the aim. Ability emotional intelligence is included in CHC theory. However, in this meta-analysis, only trait emotional intelligence is considered since, to the best of our knowledge, no ATCO selection process assesses ability emotional intelligence. Other non-cognitive predictors encompass motivation and biodata. Motivation is the proximal determinant that influences a person’s decision to allocate the effort required to perform the activities necessary to learn (Kanfer & Ackerman, 1989). Biodata refers to personal history information (Breaugh, 2009; Song, Wu, & Wang, 2019). For example, a Biographical Questionnaire (Farmer, 2002) was used to gather information on the education and background experiences of ATCOs (Pierce, Broach, Bleckley, & Byrne, 2013). Although cognitive predictors always predict the largest share of variance, non-cognitive predictors generally add incremental validity to the predictive model (O'Boyle, Humphrey, Pollack, Hawver, & Story, 2011; Zhang & Kuncel, 2020).

The purpose of this study was to conduct a psychometric meta-analysis of the articles and reports written on the predictive validity of all predictors for ATCO initial training success and on-the-job success criteria.

Methods

This meta-analysis is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement (Moher, Liberati, Tetzlaff, Altman, & Prisma Group, 2009; Page et al., 2021). We selected relevant studies by searching APA PsycINFO, Scopus, Psychology and Behavioral Sciences Collection, Web of Science Core Collection, ProQuest Dissertations and Theses and Google Scholar. The latter has been used only to ensure that no reference was missed. All relevant studies were published between December 1, 1961 and August 1, 2021 (no publication date limiter was set). The search terms and relative variants were as follows: (Air Traffic Controller$ OR Air Traffic Control Specialist$ OR ATC OR ATCS OR ATCO) and (selection OR recruitment OR validit$ OR predict$ OR success OR attrition). Search terms have been selected based on a state of the art done during the PhD thesis of the first author. No term relating to assessment/test/instrument has been used because the goal was to be as exhaustive as possible and some moderator variables are often referred by other terms. We also reviewed the references of included articles to identify relevant work. This review was done by comparing the references titles with the titles of the other articles in order to verify if there was a match. Moreover, grey literature such as institutional reports or PhD dissertations was also taken into consideration. Principal authors were contacted by email to obtain additional published or unpublished studies, only one answered and shared three unpublished studies.Principal authors contacted were the heads of the largest ATCO selection entities. We screened 12137 articles for possible inclusion (see Figure 1).

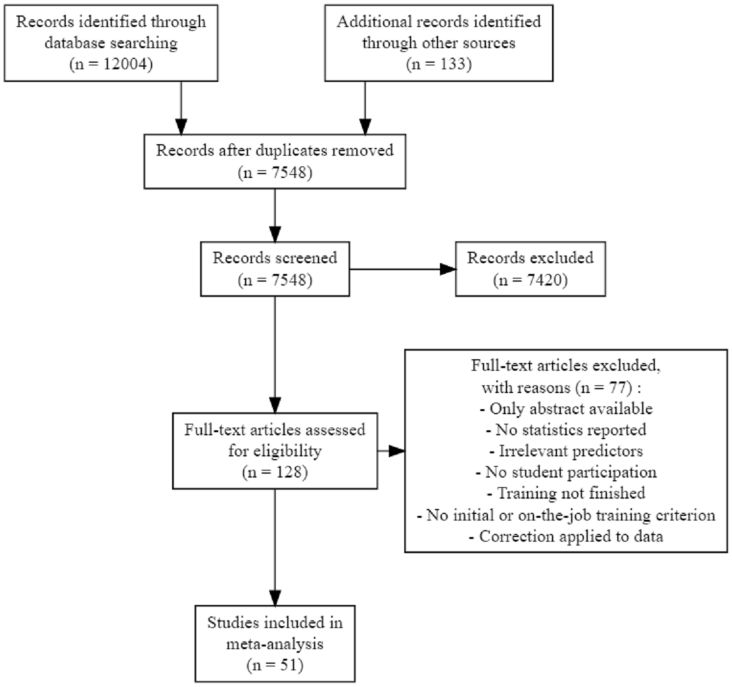

Figure 1. PRISMA flow diagram

Inclusion criteria

In several cases, multiple documents were identified which reported results from the same sample on the same variable or variables (sometimes at different time points, sometimes at the same time point), we retained the reports with the largest number of variables and the largest temporality (Wood, 2008). When studies were written initially as institutional reports and published afterward in a scientific review, the latter version was selected. In order to follow PRISMA Statement, data from previous meta-analyses were not automatically included in the current study but were included if the original documents appeared in literature searches. After duplicates were removed, 7548 articles were screened. Of these articles, 7420 were excluded based on a review of the title, of the abstract and of the journal, leaving 128 eligible articles. Because of the exhaustive keywords, many false positives were present and quickly eliminated based on journal names such as articles published in biology journals. Seventy-seven were excluded based on a full-text review of the articles leaving 51 articles for the meta-analysis (see Figure 1). To be eligible for inclusion in this meta-analytic synthesis, a study had to meet the following criteria: (1) use of cognitive, personality, motivation, or biodata predictor; (2) participants were ab-initio students; (3) use of initial or on-the-job training criterion; (4) training fully completed; (5) the article provided sample size and effect size information or sufficient statistics to calculate an effect size and (6) no correction applied to coefficients.

Statistical integration

Raw correlation coefficients (r) were extracted, r-to-z transformation was not used since the negative bias in average r is always less than the positive bias in average z (Schmidt & Hunter, 2015). If a study contained both zero-order and covariate-adjusted effect sizes, only the zero-order effect size was selected as it is most comparable with the majority of the effects. If a study contained only covariate-adjusted effect sizes, this study was excluded. If necessary, coefficients were reverse-scored so that higher scores represented higher levels of training success.

When one sample provided effect sizes for multiple outcomes relevant to one criterion, linear composites were computed since psychometric meta-analysis assumes that each effect size for a particular construct relationship comes from an independent sample (Schmidt & Hunter, 2015). When only a sample size range was given for multiple variables, the lowest sample size was retained Finally, if no correlation coefficient was given but metrics that could be transformed into coefficients were given, the transformations were performed.

Two psychometric meta-analyses were conducted with a random-effects model fitted to the data in order to draw conclusions that can be generalized beyond the specific set of studies included (Kisamore & Brannick, 2008): one meta-analysis for cognitive predictors and another for non-cognitive predictors. These two types of predictor are so different that combining them in the same meta-analysis would be unsuitable.

Random-effects variance was estimated using the Hunter-Schmidt estimator (Schmidt &

Hunter, 2015). Modeling options were kept by default: correlations were weighted by sample size in order not to introduce dependency between study weights and study effect sizes (Dahlke & Wiernik, 2019). In accordance with Gignac and Szodorai (2016), the following intervals were used: [.10, .20[for small effect, [.20, .30[for medium or typical effect and [.30, 1] as large effect. Indeed, the authors argued that these intervals were more suited for individual difference studies than Cohen's guidelines for effect sizes (Cohen, 2013).

Correction for artifactual errors

In order to reduce attenuation and to estimate the true correlation between variables, corrections for artifactual errors were applied. These methods control the impact of artifactual errors that cause the underestimation of the effect size and the increment of variability (Wiernik & Dahlke, 2020).

The principle objective of this study was to evaluate the accuracy of cognitive and non-cognitive predictors for applied-context decisions. Thus, the operational validity was calculated: the average observed effect size was corrected for range restriction in the predictor and for measurement error in the criterion (Viswesvaran, Ones, Schmidt, Le, & Oh, 2014). Range restriction is known to present statistical issues in selection psychology. Only selected participants are included in the training process and therefore can pass or fail. Thus, test performances used for predictive analysis are only those of the best participants and variance is smaller in this sample than in the target population (Dahlke & Wiernik, 2020). In this meta-analysis, indirect and direct range-restrictions were observed and corrected. Measurement errors were only corrected for the criteria since in applied-contexts, measured scores are only available and hence "correction for reliability in predictors has no sense" (Binning & Barrett, 1989). Deduction of the true score based on interpolation of measured scores and psychometric metrics would be inappropriate in an applied-context.

Since few studies reported reliability information or selection effects, individual correction was applied on raw coefficients (Schmidt & Hunter, 2015). Indeed, several methods are available to address missing correction features: artifact-distribution or individual correction methods. In the former, studies are meta-analyzed without correction (a barebone meta-analysis), then overall results are corrected with the artifact distribution observed based on weights determined by a Taylor series approximation. However, the same artifact-correction model will be applied to all effect sizes. The individual correction method will correct each study individually. This method is more appropriate since selection process generally is different. All correction features were not available for all studies but imputation was done with the R package "psychmeta". This corrected effect size will still provide a better estimate than no correction at all (Dahlke & Wiernik, 2020; Wiernik & Dahlke, 2020). For composites formed from multiple outcomes, reliability was computed based on Mosier composite formula (Mosier, 1943). Mean observed correlations ($\bar{r}$), mean corrected correlations (ρ) and 95% confidence intervals around the mean corrected correlations (95% CI) were computed.

Heterogeneity analyses

In order to assess the heterogeneity of the effect sizes, three measures were calculated: Cochrane's Q, I^2 values and 80% Credibility Intervals (Higgins, 2003). A statistically significant Q value means that the effect sizes are not homogeneous. I^2 represents the proportion of the variance across studies that is due to variation in real effects rather than sampling error (Borenstein, Higgins, Hedges, & Rothstein, 2017). I^2 is categorized at 25%, 50% and 75% as a low, moderate, and high proportion of true heterogeneity (Higgins, 2003). 80% Credibility Intervals (80% CV) were also used. Credibility intervals greater than 0.11 or inclusion of 0 suggest the presence of moderators (Koslowsky & Sagie, 1993).

Moderator analysis

Moderating variables are "variables that cause differences in the correlation between two other variables" (Schmidt & Hunter, 2015). One moderating variable was theoretical in aspect and two were of operational aspect.

The first moderator was the predictor assessed by the tests. For cognitive predictors, Cattel-Horn-Carrol theory (Schneider & McGrew, 2018) was considered in order to classify cognitive abilities measured. Coding was performed by two of the co-authors. On the basis of Cattel-Horn-Carrol definitions, the initial inter-coder agreement was substantial (κ = .78). After discussion, inter-coder agreement was perfect (κ = 1.00). Only eight CHC broad cognitive predictors were found on 51 articles. Moreover, composite scores and work sample tests were not categorized by following CHC since multiple cognitive abilities were assessed at the same time. ALMamari and Traynor (2019) proposed a classification scheme of composite scores. However, the description of the composite scores identified in the 51 articles was not sufficiently precise to classify them in the ALMamari and Traynor (2019) classification since all composite scores found would have been classified in one category based on descriptions. Thus, the cognitive meta-analysis included ten cognitive predictors: quantitative knowledge, processing speed, work sample, short-term working memory, visuo-spatial processing, cognitive composite, comprehension-knowledge, fluid reasoning, domain-specific knowledge and reading and writing. Definitions of cognitive predictors and test example are displayed on Table 1. For the non-cognitive predictors, Big Five model (Costa & McCrae, 1999) was used for personality traits. Motivation and biodata were also categorized separately, both were assessed by self-reported questionnaires or interviews. Motivation was assessed by questions such as "How did you prepare for the selection process?". Biodata has been divided into four variables: Education, Aeronautical Experience, Age and (Other) Experience. Education refers to level of education, aeronautical experience to piloting qualifications and other experience to having previous work experience. Non-cognitive composite scores were also present: several variables were averaged, these variables could belong to the category of personality traits, biodata or motivation. Each composite was computed differently.

Table 1

Definitions of cognitive predictors

Predictor	Definition	Example
Quantitative knowledge	The depth and breadth of declarative and procedural knowledge related to mathematics.	ATSAT - Applied-math (Kelley, 2012)
Processing speed	The ability to control attention to automatically, quickly and fluently perform relatively simple repetitive cognitive tasks. Attentional fluency or attentional speediness.	Moran Repetitive Measurements - Perceptual speed (Cobb, 1964)
Work sample	A work sample test is a test in which the applicant performs a selected set of actual tasks that are physically and/or psychologically similar to those performed on the job (Ployhart et al., 2005).	MCAT (Della Rocco & Mannibg, 1990)
Short-term working memory	The ability to maintain and manipulate information in active attention. The mind's mental "scratchpad" or "workbench".	CTMM-coins (Cobb, 1971)
Visuo-spatial processing	The ability to make use of simulated mental imagery to solve problems. Perceiving, discriminating and manipulating images in the "mind's eye".	Directional Heading Test (Cobb & Mathews, 1972)
Cognitive composite	In situations where multiple cognitive tests are administered, scores from individual tests are frequently combined to produce a composite score (Song, Lin, Ward, & Fine, 2013).	ASVAB composites (Carretta & King, 2008)
Comprehension-knowledge	The ability to comprehend and communicate culturally-valued knowledge. Gc includes the depth and breadth of both declarative and procedural knowledgeand skills such as language, words, and general knowledge developed through experience, learning and acculturation.	SLS vocabulary (Della Rocco et al., 1992)
Fluid reasoning	The use of deliberate and controlled procedures (often requiring focused attention) to solve novel "on the spot" problems that cannot be solved by using previously learned habits, schemas, and scripts.	CTMM analogies (Cobb et al., 1968)
Domain-specific knowledge	The depth, breadth, and mastery of specialized declarative and procedural knowledge (knowledge not all members of a society are expected to have).	OKT (Rock et al., 1984)
Reading and writing	The depth and breadth of declarative and procedural knowledge and skills related to written language.	ASVAB - Paragraph Comprehension (Carretta & King, 2008)

Note. Unless otherwise noted, all definitions are from Scheneider and McGrew (2018).

The second moderator was the period of publication and evolution of air traffic operations. Since this meta-analysis covered articles from 1960 to today, the timespan was divided into three periods of 20 years (1960 to 1980, 1980 to 2000 and 2000 to 2020; a division confirmed by data distribution).

The third moderator was the nature of criterion used: initial training (mainly theoretical and academic) or on-the-job training (mainly practical and operational).

Initially, other moderators (control position and success rate) were considered but too few articles gave clear information on these subjects.

Moderators were analyzed with a subgrouping approach since they were categorical (Schmidt, 2017). A moderator modality was discarded if only one study has been published on it. A one-way analysis of variance (ANOVA) was performed for each moderator. To check for any statistically significant difference between moderator modalities, a maximum *p*-value of .05 was set. Post-hoc analyse were based on Wald-type pairwise comparisons for each level of categorical moderators. Corrections were performed using Benjamini-Hochberg method (False Discovery Rate), this method being less severe and more powerful than Bonferroni correction when there are multiple significant effects (Benjamini & Hochberg, 1995).

Publication bias

Publication bias is "the term for what occurs whenever the research that appears in the published literature is systematically unrepresentative of the population of completed studies" (Rothstein, Sutton, & Borenstein, 2005). In psychology and especially in selection psychology, the non-publication of non-significant results cannot be ruled out. Although a substantial effort have been made to include as many published and unpublished studies as possible, the average estimates could be affected by publication bias. Publication bias was checked by several means. First of all, each document was coded according to its publication status, i.e. peer-reviewed articles or non peer-reviewed documents (PhD dissertation and institutional report). This variable was used as a moderator in order to approximate the impact of publication bias by comparing the mean effect sizes for peer-reviewed documents and documents without peer-review. In addition, publication bias were tested by precision-effect test (PET) and precision-effect estimate with standard error (PEESE) for each meta-analysis (Carter, Schönbrodt, Gervais, & Hilgard, 2019; Wiernik & Dahlke, 2020).

Outlier identification

Raw coefficients should be screened for outliers since extreme coefficients may bias meta-analytic estimates (Viechtbauer & Cheung, 2010). Cook's distances and externally standardized residuals were used. Two studies were identified based on these metrics (Broach, 2019; Dean & Broach, 2012) but they were kept, as suggested by Carter et al. (2019), since results were different when these two studies were excluded.

R packages

Under R (R Core Team, 2020), the package **esc** (Lüdecke, 2019) was used for effect size transformation, the package **psychmeta** (Dahlke & Wiernik, 2019) for psychometric meta-analytic computation, the package **metafor** (Viechtbauer, 2010) for outlier analysis and the package **papaja** (Aust & Barth, 2020) for redaction.

Results

Data description

The meta-analysis is composed of 51 studies with a total of 65839 participants and 651 correlations. The settings represented in the study were essentially American: Federal Aviation Administration (36), US Navy (6), US Air Force (4), DFS (3), FEAST (3), US Marine Corps (2), Royal Netherlands Air Force (2) and New Zealand Airways (1). Thirteen studies were published between 1960 and 1980 and provided 231 coefficients, 19 between 1980 and 2000 with 208 coefficients and 20 between 2000 and 2020 with 212 coefficients. The majority of the studies were institutional reports (28), 19 were publications in academic journals and four were PhD dissertations. Many studies reported only cognitive predictors (27), followed by both cognitive and non-cognitive predictors (18) and non-cognitive predictors only (6). The majority of predictors were cognitive (424).

Cognitive meta-analysis

Overall effect. In the cognitive meta-analysis, there were 45 studies included with an overall corrected effect [1] of $r = .37$ with a 95% confidence interval ranging from .31 to .42.

Publication bias. No significant difference was measured between peer-reviewed documents and documents without peer-review ($F(1,40.5) = 1.6$; $p = .21$). PET and PEESE analyses (see Figure 2) show significant relation of effect size with either standard error or sampling error variance only for barebone meta-analysis ($\beta = 2.37$, 95% CI [0.49, 4.24] and $\beta = 23.63$, 95% CI [8.80, 38.45]). For corrected meta-analysis, PET and PEESE analyses (see Figure 2) show non-significant relation of effect size with either standard error or sampling error variance ($\beta = 0.56$, 95% CI [-2.34, 3.46] and $\beta = 1.89$, 95% CI [-11.39, 15.16]). Therefore, results of this meta-analysis need to be treated with caution since a publication bias may exist.

[1] For clarity, only individually corrected effect size will be discussed.

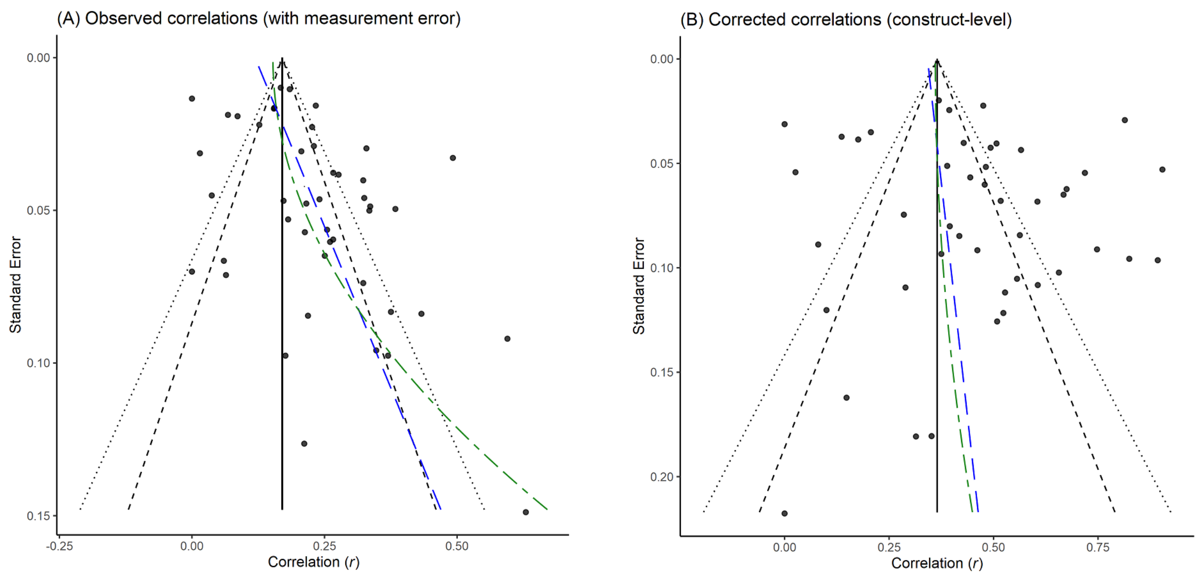

Figure 2. PET and PEESE plot for cognitive meta-analysis

Heterogeneity. Cochrane's Q was significant: $Q(44) = 688.657$, $p < .001$. I^2 was 93.13%. Together, these metrics suggest the existence of substantive moderators of the cognitive predictor-training success relationship. Moderator analysis was therefore totally relevant.

Moderator analyses. Table 2 contains all moderator analyses conducted with the corresponding number of effect sizes (k), sample size (N), observed and corrected effect sizes ($\bar{r}$ and ρ) with their respective standard deviation, the 95% lower and upper confidence intervals and the 80% lower and upper credibility intervals.

Table 2

Meta-analytic estimates of effects of cognitive predictors on training success

Moderator	k	N	$\overline{r}$	SD_r	$\overline{\rho}$	SD_ρ	95% CI	80% CR
Overall	45	57189	.17	.1	.37	.19	[.31, .42]	[.12, .61]
Predictor								
Quantitative knowledge	6	2 458	.25	.05	.40	.10	[.28, .52]	[.26, .54]
Processing speed	15	7 785	.22	.07	.42	.12	[.34, .50]	[.26, .58]
Work sample	18	13820	.24	.11	.41	.22	[.28, .52]	[.10, .71]
Short-term working memory	8	1 851	.21	.10	.39	.13	[.26, .53]	[.22, .57]
Visuo-spatial processing	14	9 567	.18	.09	.41	.11	[.33, .48]	[.26, .55]
Cognitive composite	34	52291	.17	.09	.37	.19	[.30, .43]	[.12, .62]
Comprehension-knowledge	3	1 523	.09	.03	.23	.00	[.10, .37]	[.23, .23]
Fluid reasoning	11	11 510	.10	.10	.19	.19	[.06, .32]	[-.07, .45]
Domain-specific knowledge	9	9 547	.14	.10	.19	.12	[.10, .29]	[.03, .36]
Reading and writing	3	1 012	.19	.05	.36	.19	[−.15, .87]	[.01, .71]
Period of publication								
1960-1980	11	5 292	.28	.08	.50	.15	[.39, .61]	[.29, .71]
1980-2000	17	42 046	.16	.08	.36	.16	[.27, .44]	[.14, .58]
2000-2020	18	10502	.17	.13	.32	.24	[.20, .44]	[.00, .64]
Criterion nature								
Initial training	30	41744	.23	.08	.50	.10	[.46, .54]	[.36, .64]
On-the-job training	25	34532	.08	.07	.18	.14	[.12, .24]	[.00, .37]

Note. k = number of studies contributing to meta-analysis; N = total sample size; $\overline{r}$ = mean observed correlation; SD_r = observed standard deviation of r; p = mean true-score correlation; SD_ρ = observed standard deviation of $\overline{\rho}$; CI = confidence interval around $\overline{\rho}$; CR = credibility interval around $\overline{\rho}$. Correlations corrected individually.

Moderation by predictor. The relations with the criterion were significant for all predictors except one: reading and writing (r = .36 ; 95% CI [-.15, .87]). Quantitative knowledge correlated with the criterion (r = 40 ; 95% CI [.28, .52]), as well as processing speed (r = .42 ; 95% CI [.34, .49]), work sample (r = .41 ; 95% CI [.29, .52]), short-term working memory (r = .39 ; 95% CI [.26, .52]), visuo-spatial processing (r = .41 ; 95% CI [.33, .48]) and cognitive composite (r = .37 ; 95% CI [.30, .44]). An ANOVA highlighted a main effect of the predictor factor (F(9,22.2) = 3.89; p = .004). A forest plot of predictor factors is displayed in Figure 3.

Quantitative knowledge predicts significantly better than comprehension-knowledge (95% CI [.03, .30]), fluid reasoning (95% CI [.05, .67]) and domain-specific knowledge (95% CI [.07, .34]). Analogously, processing speed predicts significantly better than comprehension-knowledge (95% CI [.08, .28]), fluid reasoning (95% CI [.09, .37]) and domain-specific knowledge (95% CI [.11, .34]). Work sample predicts better than comprehension-knowledge (95% CI [.04, .31]), domain-specific knowledge (95% CI [.07, .36]) and fluid reasoning (95% CI [.05, .39]). Short-term working memory predicts better than comprehension-knowledge (95% CI [.01, .31]), domain-specific knowledge (95% CI [.05, .35]) and fluid reasoning (95% CI [.03, .38]). Lastly, visuo-spatial processing predicts significantly better than domain-specific knowledge (95% CI [.07, .27]), domain-specific knowledge (95% CI [.10, .32]) and fluid reasoning (95% CI [.08, .36]).

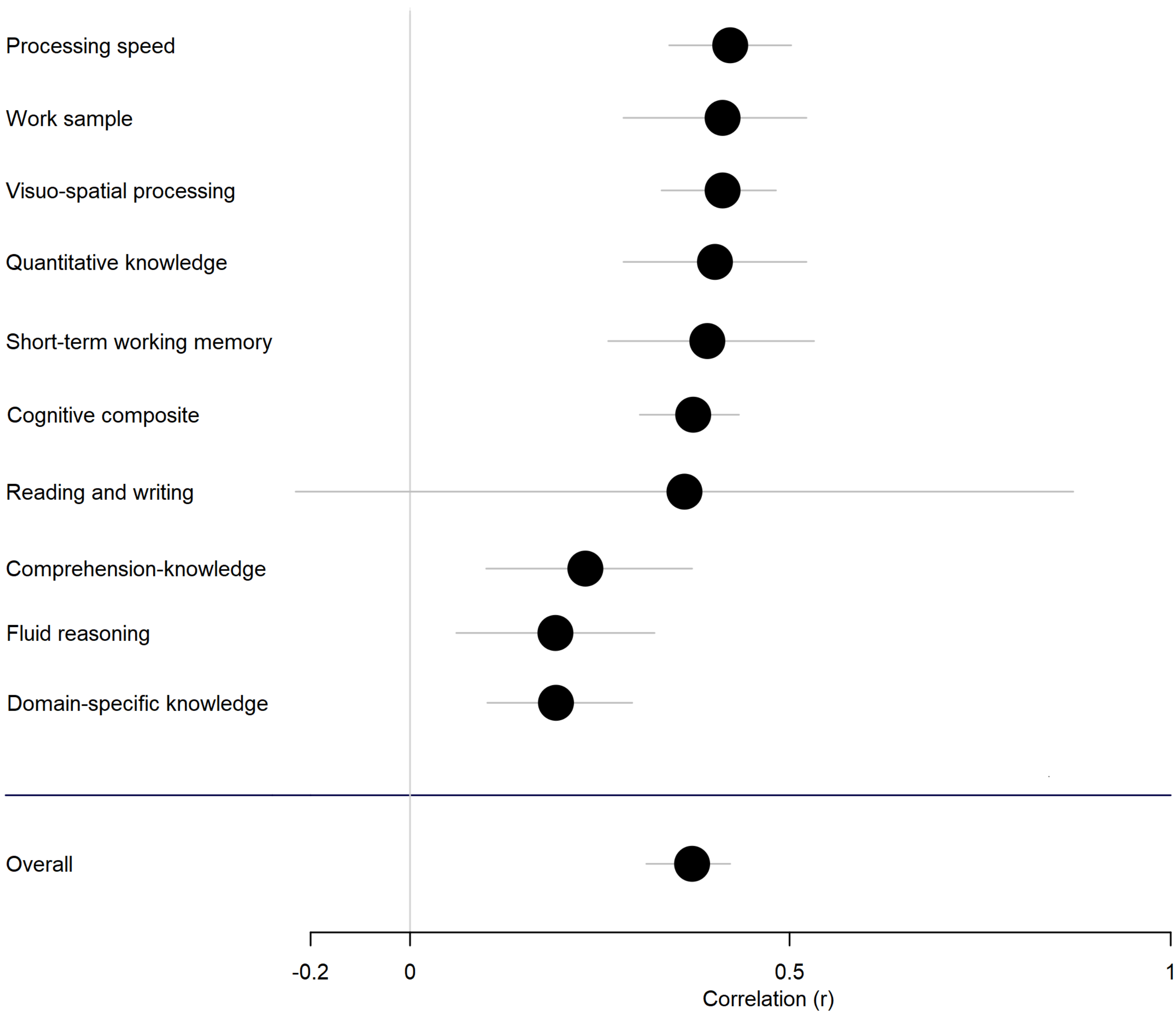

Figure 3. Forest plot with only cognitive predictors

Moderation by period of publication. The associations with the criterion were significant for the three periods: all studies between 1960 and 1980 correlated with the criterion (r = .50 ; 95% CI [.39, .61]) as well as between 1980 and 2000 (r = .36 ; 95% CI [.27, .44]) and also between 2000 and 2020 (r = .32 ; 95% CI [.20, .44]). A main effect of this moderator was observed (F(2,26.5) = 3.48 ; p = .045). Post-Hoc analyses revealed that studies published between 1960 and 1980 had higher validity than studies between 1980 and 2000 (95% CI [-.28, -.01]) and than studies between 2000 and 2020 (95% CI [-.34, -.02]).

Moderation by criterion nature. The associations between all cognitive predictors and both criterion natures were significant: initial training (r = .50 ; 95% CI [.46, .54]) and on-the-job training (r = .18 ; 95% CI [.12, .24]). Initial training is more precisely predicted than on-the-job training (95% CI [.24, .29]). Based on this significant difference, a supplementary analysis was thus conducted to explore the interaction between each cognitive predictor [2] and criterion nature. A forest plot of this interaction is displayed in Figure 4.

Cognitive composite better predicted (95% CI [.23, .41]) initial training criterion (r = .52 ; 95% CI [.47, .57]) than on-the-job training criterion (r = .20 , 95% CI [.12, .27]). The same pattern was measured for the work sample predictor (95% CI [.14, .47]): initial training criterion (r = .56 ; 95% CI [.43, .69]) vs on-the-job training criterion (r = .26 ; 95% CI [.13, .39]) and for fluid reasoning (95% CI [.09, .60]): initial training criterion (r = .33 ; 95% CI [.26, .39]) vs on-the-job training criterion (r = -.02 ; 95% CI [-.49, .46]). Others predictors were not moderated by criterion criterion.

[2]Comprehension-knowledge predictor was discarded since no study used it for initial training predictivity.

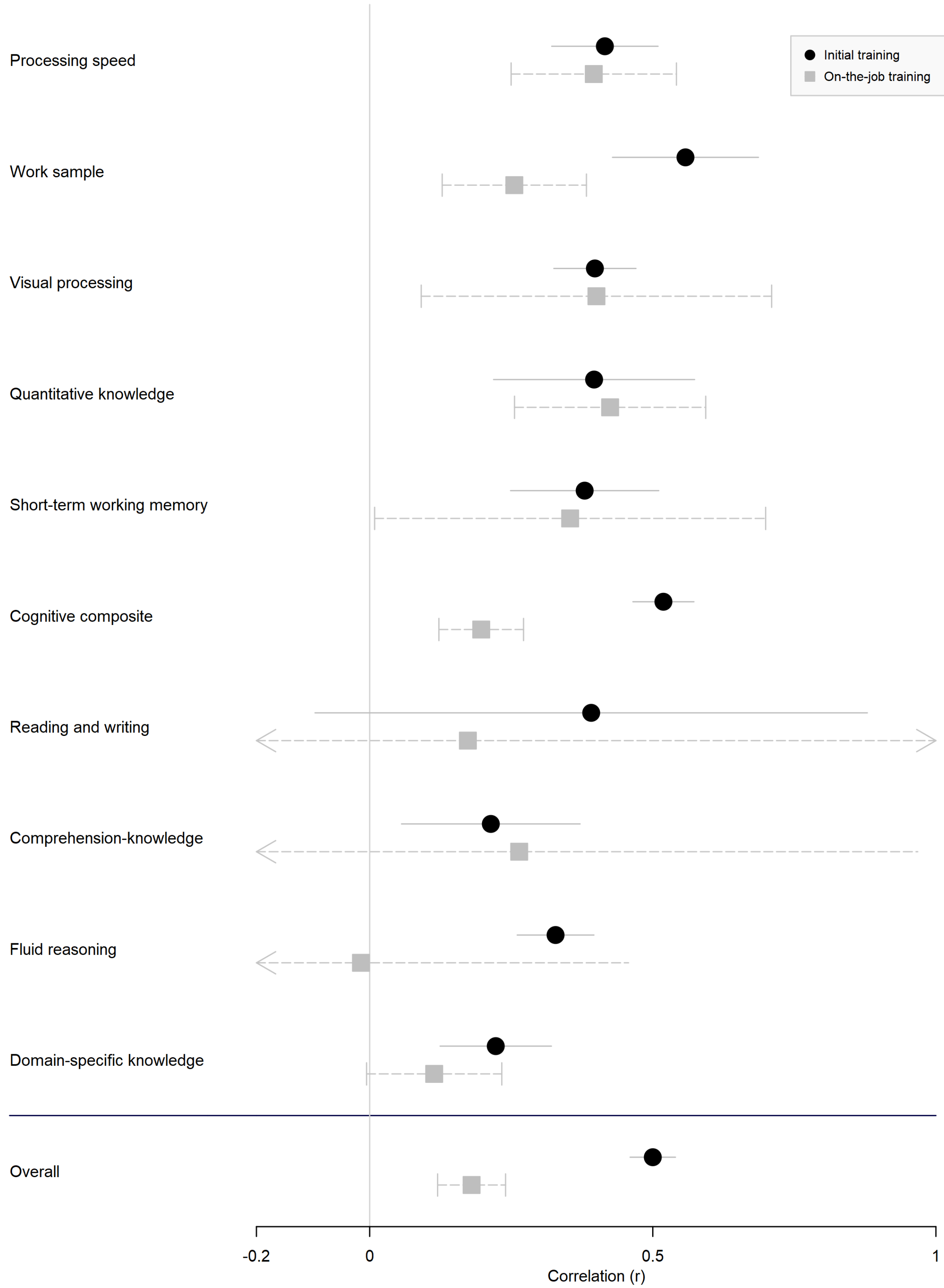

Figure 4. Forest plot with only cognitive predictors by criterion nature

Non-Cognitive meta-analysis

Overall effect. In the non-cognitive meta-analysis, there were 24 studies included with an overall corrected effect of $r = .15$ ($df = 23$, $p < .001$) with a 95% confidence interval ranging from .08 to .22.

Publication bias. Moderator analysis highlighted a significant difference between peer-reviewed documents and documents without peer-review ($F(1,22) = 5.24$; $p = .03$) with results published in peer-reviewed documents lower than in documents without peer-review. PET and PEESE analyses (see Figure 5) show negligible relation of effect size with either standard error (for barebone meta-analysis : $\beta = 2.17$, 95% CI [-0.79, 5.14], for corrected meta-analysis : $\beta = 1.98$, 95% CI [-1.21, 5.17]) or sampling error variance (for barebone meta-analysis : $\beta = 21.62$, 95% CI [-0.37, 43.60], for corrected meta-analysis : $\beta = 15.67$, 95% CI [-5.30, 36.63]). Therefore, results of this meta-analysis need to be treated with caution since a publication bias exists.

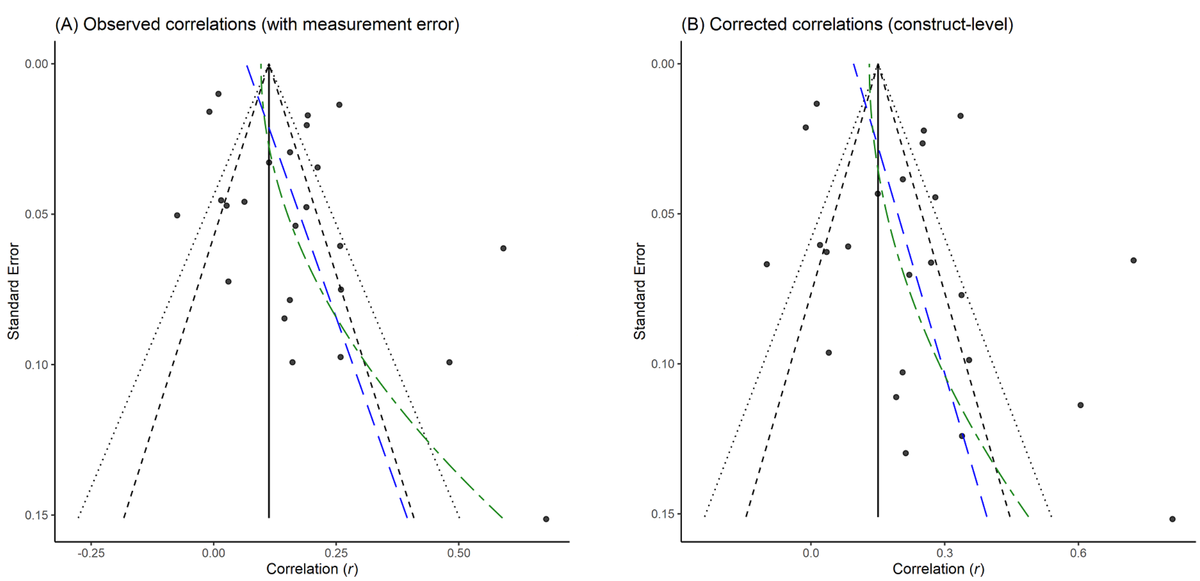

Figure 5. PET and PEESE plot for non-cognitive meta-analysis

Heterogeneity. Cochrane's Q was significant: $Q(23) = 464.571$, $p < .001$. I^2 was 95.05% and 80% CV overlaps with zero. Based on these three metrics, we can say there are significantly different effect sizes across the studies. Moderator analysis was therefore relevant.

Moderator analyses. Table 3 contains all moderator analyses conducted with the corresponding number of effect sizes (k), sample size (N), observed and corrected effect sizes ($\overline{r}$ and ρ) with their respective standard deviation, the 95% lower and upper confidence intervals and the 80% lower and upper credibility intervals.

Table 3

Meta-analytic estimates of effects of non-cognitive predictors on training success

Moderator	k	N	$\overline{r}$	SD_r	$\overline{\rho}$	SD_ρ	95% CI	80% CR
Overall	24	31357	.11	.12	.15	.16	[.08, .22]	[−.05, .35]
Predictor								
Non-cognitive composite	5	4167	.34	.11	.44	.13	[.27, .60]	[.24, .63]
Education	7	13303	.25	.17	.33	.21	[.13, .52]	[.03, .63]
Aeronautical Experience	5	5 336	.13	.09	.17	.11	[.04, .31]	[.01, .33]
Motivation	8	10 256	.12	.08	.16	.10	[.07, .24]	[.02, .29]
Experience	2	474	.12	.02	.16	.02	[−.06, .39]	[.16, .16]
Age	4	4340	.19	.13	.24	.16	[−.02, .50]	[−.02, .50]
Agreeableness	4	10 807	.01	.01	.02	.02	[−.01, .04]	[.02, .02]
Conscientiousness	6	11 385	.01	.03	.01	.05	[−.03, .06]	[−.04, .06]
Extraversion	5	15 351	.01	.02	.01	.03	[−.03, .05]	[−.02, .04]
Neuroticism	5	11 281	.01	.02	.02	.02	[−.01, .04]	[.02, .02]
Openness to experience	5	10 911	.01	.04	.01	.04	[−.04, .06]	[−.05, .06]
Period of publication								
1960-1980	4	3 954	.16	.09	.21	.11	[.02, .40]	[.02, .39]
1980-2000	9	23 335	.09	.12	.12	.16	[.00, .25]	[−.10, .34]
2000-2020	11	4067	.19	.14	.25	.17	[.13, .37]	[.03, .48]
Criterion nature								
Initial training	16	22438	.07	.09	.09	.12	[.02, .16]	[−.07, .25]
On-the-job training	13	21642	.12	.12	.16	.17	[.06, .26]	[−.07, .38]

Note. k = number of studies contributing to meta-analysis; N = total sample size; $\overline{r}$ = mean observed correlation; SD_r = observed standard deviation of r; p = mean true-score correlation; SD_ρ = observed standard deviation of $\overline{\rho}$; CI = confidence interval around $\overline{\rho}$; CR = credibility interval around $\overline{\rho}$. Correlations corrected individually.

Moderation by predictor. At a lower theoretical level, the relations between the predictor and the criterion were non significant except for four: non-cognitive composite (r = .44 ; 95% CI [.27, .60]), education (r = .33 ; 95% CI [.13, .52]), then aeronautical experience (r = .17 ; 95% CI [.04, .31]) and motivation (r = .16 ; 95% CI [.07, .24]).

An ANOVA highlighted a main effect of the predictor factor (F(10,14.1) = 10.6 ; p = .001). A forest plot of predictor factor is displayed in Figure 6.

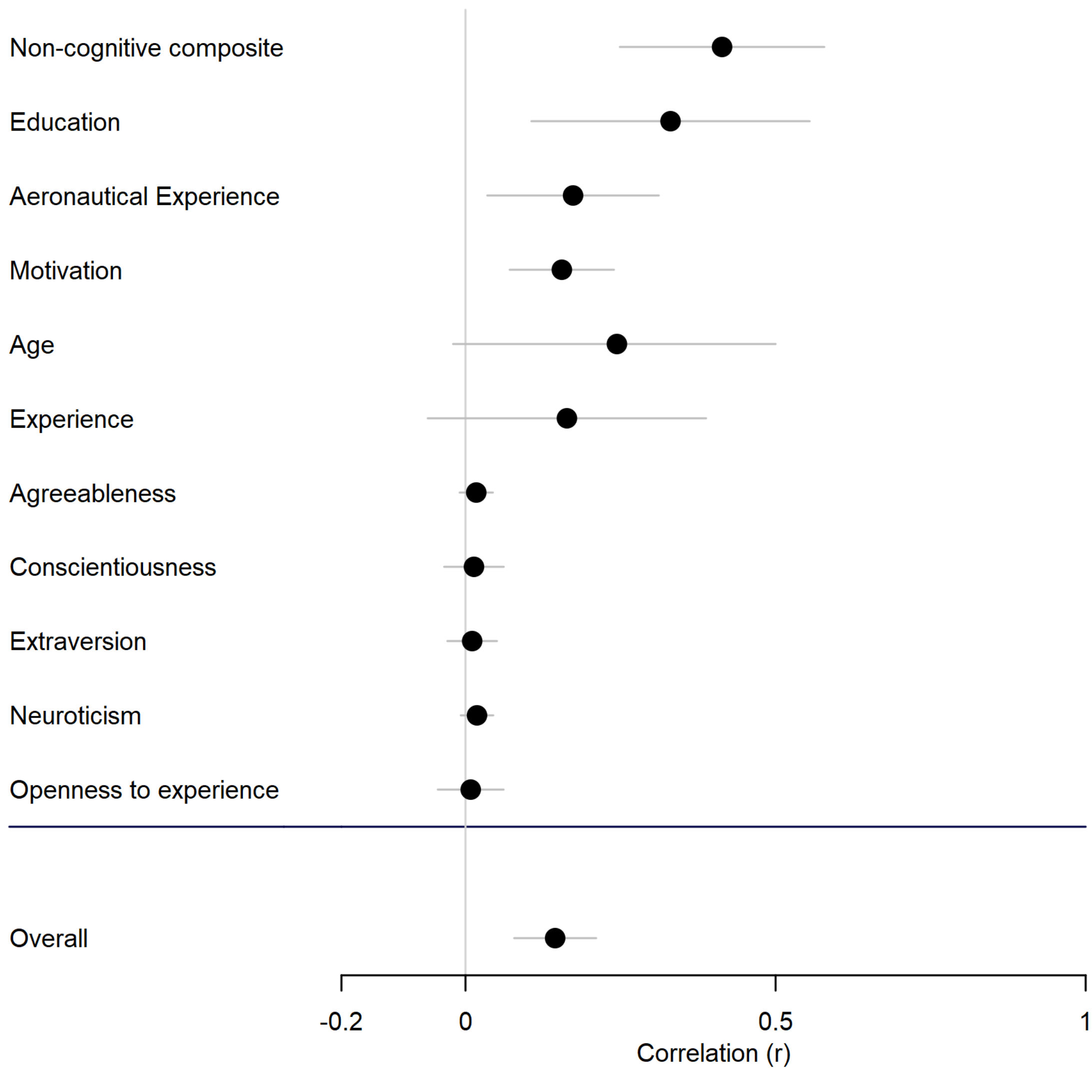

Figure 6. Forest plot with only non-cognitive predictors

Non-cognitive composite was significantly better than each Big Five personality trait: agreeableness (95% CI [.28, .56]), conscientiousness (95% CI [.28, .56]), extraversion (95% CI [.29, .56]), neuroticism (95% CI [.28 ; .55]) and openness to experience (95% CI [.28, .57]). The same results were obtained for education, aeronautical experience and motivation: all three were significantly better than the Big Five personality traits.

Moderation by period of publication. The associations with the criterion were significant for the three periods: all studies between 1960 and 1980 correlated with the criterion ($r = .21$; 95% CI [.02, .40]) as well as between 1980 and 2000 ($r = .12$; 95% CI [.00, .25]) and also between 2000 and 2020 ($r = .25$; 95% CI [.13, .37]). No significant difference was measured ($F(2,10.2) = 1.41$; $p = 0.29$).

Moderation by criterion nature. The associations between all non-cognitive predictors and both criterion natures were significant: initial training ($r = .09$; 95% CI [.03, .16]) and on-the-job training ($r = .16$; 95% CI [.06, .26]). The two associations were not significantly different from each other ($F(1,21.7) = 1.27$; $p = .27$), no supplementary analysis was performed.

Discussion

The main goal of this meta-analysis was to examine the relationship between predictors and ATCO training success through meta-analyses comparing the diverse predictors and instruments used. The meta-analysis covered 51 studies since 1960 with a total of 65839 participants, 651 correlations were included. These studies employed a large variety of predictors encompassing both cognitive and non-cognitive predictors. However, some predictors were used despite a lack of validity. Contrary to popular belief in aviation psychology, our findings suggested that personality traits do not play a role in success for air traffic controller training. Compared to Martinussen et al. (2000), results are almost similar: the effect sizes are of the same importance and personality is not significant. The results will be also discussed in relation to pilot selection with cognitive (Johnson, Barron, Carretta, & Rose, 2017) and non-cognitive predictors (Hunter & Burke, 1994). Recently, Sackett, Zhang, Berry, and Lievens (2021) used new corrections methods on old datasets and observed reductions in validity, results will be also compared against this. In a general way, our results are a little inflated compared to Sackett et al. (2021). Based on their re-analysis of Schmidt and Hunter (1998), cognitive abilities decreased from .51 to .33 and work sample from .54 to .33. However, ranking of predictors is the same in both analyses : work sample and cognitive abilities have a better predictive power than non-cognitive abilities. A recent meta-analysis (Speer et al., 2021) focused on biodata, our results will be compared.

Cognitive meta-analysis

In the cognitive meta-analysis, the overall corrected effect is $r = .37$. The cognitive overall corrected effect can be interpreted as large.

Top 6: Quantitative knowledge, processing speed, work sample, short-term working memory, visuo-spatial processing, cognitive composite. Six predictors can be distinguished by their large predictive validity: quantitative knowledge ($r = .40$; 95% CI

[.28, .52]), processing speed (r = .42 ; 95% CI [.34, .49]), work sample (r = .41 ; 95% CI [.29, .52]), short-term working memory (r = .39 ; 95% CI [.26, .52]), visuo-spatial processing (r = .41 ; 95% CI [.33, .48]) and cognitive composite (r = .37 ; 95% CI [.30, .44]). These six predictors have a large effect ($r > .30$).

Quantitative knowledge is "the depth and breadth of declarative and procedural knowledge related to mathematics" (Schneider & McGrew, 2018). This ability is known to be linked to general decision making skill: both require practical probabilistic reasoning and skilled metacognition (e.g., accurately evaluating and integrating thoughts, feelings, risks, and values) (Cokely, 2018; Huck, 2020), a highly important skill for an ATCO. This could explain the large effect (r = .40). In Martinussen et al. (2000), quantitative knowledge was categorized as mathematical ability and was the fourth best predictor (r = .25). In pilot selection meta-analysis (Johnson et al., 2017), quantitative knowledge was the second best predictor for all criterion combined (r = .23). It was the best for academic grades (r = .33).

Information processing is critical for learning information, including new knowledge and skills (Kanfer & Ackerman, 2005). Two broad cognitive abilities are, by extension, critical for information processing: processing speed and short-term memory. The higher the processing speed, the faster the learning. Large quantities of information need to be processed by an ATCO in real time and fast processing means that more information can be processed at the same time. This could explain the large effect (r = .42). Processing speed was not incorporated in Martinussen et al. (2000) but was in Johnson et al. (2017) meta-analysis. Processing speed, categorized as perceptual speed, was the best predictor for pilot success (r = .24).

Short-term working memory broad ability encompasses both working memory capacity and attentional control. Capacity is the number of items stored in working memory and attentional control helps not to be distracted by information that can be irrelevant to the task and to focus on information relevant. Processing speed and short term working memory appear to be interrelated since they could be limited by similar features of the neurocognitive system (Dang, Braeken, Colom, Ferrer, & Liu, 2015). Neural speed of information processing explains the same percentage of variance in intelligence as working memory (Frischkorn, Schubert, & Hagemann, 2019). The large effect measured (r = .39) can be considered as equal to processing speed (r = .44), this equality is in line with Dang et al. (2015). A short-term working memory predictor was not included in Martinussen et al. (2000) nor in Johnson et al. (2017).

Redding, Cannon, and Seamster (1992) proposed that "expertise in ATC appears to be especially significant attributes of expertise in complex, time-constrained, multi-tasking environments". Work sample tests are complex, time-constrained, multi-tasking environments. Therefore, a candidate with a high score in work sample tests can be considered as a high performer in multi-tasking, an element of importance in ATC (Redding et al., 1992). The large effect (r = .41) supports this idea. In Martinussen et al. (2000), work sample was the best predictor (r = .30) In order to be a high performer in multi-tasking, processing speed should be fast and working memory high. Working memory was found to be a predictor of multi-tasking performance (Bühner, König, Pick, & Krumm, 2006). Our results is in line with Sackett et al. (2021), they obtained a large effect with their new correction method.

Based on job analyses (Broach, 2013; Morath, Quartetti, Bayless, & Archambault, 2001), visuo-spatial processing is known as critical for ATCO. Visuo-spatial processing is "the ability to make use of simulated mental imagery to solve problems" (Schneider & McGrew, 2018). The large effect (r = .41) supports this criticality. According to Schneider and McGrew (2012), visualization is generally the narrow cognitive ability assessed by visuo-spatial processing tests while other narrow abilities exist in the visuo-spatial processing category. For the past two decades, visuo-spatial processing has gained in significance, in particular in STEM learning and prediction of success (Buckley, Seery, & Canty, 2018; Lubinski, 2010). The same results are observed in pilots (ALMamari & Traynor, 2020). In Martinussen et al. (2000), visuo-spatial processing, known as spatial ability, was the third best predictor (r = .28). For pilots (Johnson et al., 2017), visuo-spatial processing has a medium effect (r = 18).

Cognitive composite is a sort of black box (Bobko et al., 2007), multiple test scores are combined together. Interpretation of the large effect (r = .37) is difficult since multiple and different cognitive abilities can be assessed in the same time. For ALMamari and Traynor (2019), "it is a combination of skills and competencies that jointly contribute to the scores given to applicants and influence organization's employment decisions".

Large moderation effect by criterion nature. For cognitive predictors, an important moderator is criterion nature. Almost all predictors have larger predictive validities for initial training than for on-the-job training. On-the-job training requires all the knowledge and skills obtained in initial training but as it is in real situation, ATCOs have also to work in collaboration. Social competencies and emotional intelligence are not learned in initial training whereas these competencies are necessary during on-the-job training and could explain training failure.

Composite scores were mediated by criterion nature. Prediction was very high (r = .52) for initial training but low for on-the-job training (r = .20). Initial training is a phase where many competencies and much knowledge need to be acquired in an ATCO school. These composite scores could represent a g factor even if the number of tests, contents and weights were scarcely specified in the articles or were very different from one study to another (Major, Johnson, & Bouchard Jr, 2011). A low-performer student during initial training should have a low composite score and will not pass in on-the-job training, hence the drop in predictive validity. A medium-to-large effect was also measured between composite score and training criteria in Martinussen et al. (2000). Work sample (r = .56 vs r = .26) and fluid reasoning (r = .33 vs r = .0) predictors show the same pattern.

The comprehension-knowledge predictor shows a small-to-medium effect (r = .23), as in Martinussen et al. (2000). Since only three studies tested this predictor and none with initial training criterion, more studies are necessary. For the reading and writing predictor, a non-significant correlation was observed (r = .36 ; 95% CI [-.15, .87]). There are too few studies and too much variability between them, thus more studies are also necessary.

Small moderation by time. Studies published between 1960 and 1980 showed better results than studies published after this period. This could be due to the knowledge of the tests by the candidates. Using Internet, candidates can exchange about selection processes more easily nowadays than in the 1960s when no such community existed. As a result of this knowledge, the private sector has seized upon this opportunity and offers coaching. When candidates prepare for the selection process, their performances are higher than usual. However, test-taking preparation is often test-specific (Matton, Vautier, & Raufaste, 2009) and is not associated to an increase in corresponding broad ability.

Non-Cognitive meta-analysis

The non-cognitive overall corrected effect is small (r = .15). A cautionary note needs to be added: the majority of articles using non-cognitive predictors did not control for the effects of cognitive abilities in their predictive analyses.

Top 2: Non-cognitive composite and education. Two non-cognitive predictors show a large effect: non-cognitive composite and education. Non-cognitive composite is the best predictor (r = .44). As for cognitive composite, contents and weights were scarcely specified or were very different from one study to another. As such, motivation can be averaged with one personality trait or two or with experience. To our knowledge, no meta-analysis has studied non-cognitive composite but biodata composites were included in Speer et al. (2021). Non-cognitive composite can be composed by biodata but others data also. Speer et al. (2021) obtained strong effects between composite scale and job performance ratings (r = .37). Our results are similar. Concerning education (r = .33), each additional year of education increases the level of cognitive ability (Falch & Sandgren Massih, 2011). In pilot selection, no relationship was measured between education and criterion (Hunter & Burke, 1994.).

Aeronautical experience and motivation. Aeronautical experience has a small effect (r = .17). Piloting experience will help a little. Motivational aspect has also a small effect (r = .16). Motivation drives cognitive process (Ganuthula & Sinha, 2019) and thus can explain the cognitive performance of a candidate (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). Motivation was assessed by self-report so could be subject to faking.

Personality traits. Personality traits are generally recognized for their predictive validity with academic performance (Mammadov, 2021) and job performance (Barrick & Mount, 1991). However no such recognition is present in this meta-analysis. All five dimensions are almost non-related with the criterion and are non-significant: Agreeableness, Conscientiousness, Extraversion, Neuroticism and Openness to experience, all overlap with zero. The same results were obtained in Martinussen et al. (2000) and Hunter and Burke (1994). Including at least ten thousand participants, the sample is sufficiently large to be reliable. Self-reporting might be the explanation: stakes are high for candidates, faking and social desirability may misrepresent the truth (Christiansen et al., 2021). Mean scores are higher and validity coefficients are lower since mean scores are false (Jeong, Christiansen, Robie, Kung, & Kinney, 2017). The link between personality predictor and success criterion may exist but the measurement of personality is not valid. Self-reporting could be replaced by situational judgment test (Schröder, Heimann, Ingold, & Kleinmann, 2021). Using other sources of information such as digital traces (Stachl et al., 2020; Wiernik et al., 2020) or a selection camp as used by Skyguide (Swiss's ANSP) (Skyguide, 2021) or moving away from the Big Five to more complex models (Mõttus et al., 2020) could improve predictive validity of personality traits. Compared to Sackett et al. (2021), our results are really low : they obtained a mean value around .20. In a second-order meta-analysis (a meta-analysis of plural meta-analyses), Barrick, Mount, and Judge (2001) showed low-to-medium effects between general training performance and extraversion (r = .28), neuroticism (r = .09), agreeableness (r = .14), conscientiousness (r = .27) and openness (r = .33). When Barrick et al. (2001) restricted their analysis on skilled professions (comparable to ATCO profession), validities between job performance (not only training performance) were strongly decreased for extraversion (r=.06), neuroticism (r = .12), agreeableness (r = .10), conscientiousness (r = .23), openness (r = .05). Conscientiousness validity has slighlty decreased between general training performance and job performance for skilled professions but non-significant results were obtained in our meta-analysis.

Age and experience predictors. Both predictors were not significant: age with r = .24 (95% CI [-.02, .50]) and experience with r = .16 (95% CI [-.06, .39]). The lack of relation could be explained by the low number of studies using these predictors. Age was not a significant predictor in pilot selection (Hunter & Burke, 1994).

Limits

The findings of the meta-analysis should be considered with caution because there were several limitations. Firstly, some predictors were employed only in a few studies. Secondly, the Cattel-Horn-Carroll theory (Schneider & McGrew, 2018) was used for the predictor classification. According to McGrew (2005), "The CHC theory of intelligence is the tent that houses the two most prominent psychometric theoretical models of human cognitive abilities" (p. 137). Based on a cross-battery CFAs of six intelligence tests and 4000 participants, Caemmerer, Keith, and Reynolds (2020) supported the Cattell-Horn-Carroll theory. However, not all researchers agree with McGrew (2005) such as Wasserman (2019). Other theories of intelligence exist and could have been used such as Process overlap theory (Kovacs & Conway, 2016, 2019). In Schmank, Goring, Kovacs, and Conway (2021), no significant difference was measured between Cattell-Horn-Carroll theory (compatible with latent variable models) and Process Overlap Theory (compatible with network models). In addition, classification of composite scores proposed by ALMamari and Traynor (2019) could be used if a full description of these scores were specified by authors. Thirdly, another limit was the application of correction in order to compensate for a restricted sample, application done not always with all necessary metrics. Predictive validity studies based on unrestricted samples delete the need for correction, as in Hoermann, Noser, and Stelling (2018). Or, in a less extreme way, all metrics should be provided in order to correct the data accurately (Sackett et al., 2021). Finally, publication bias was measured: many non-published studies appear to be missing.

Pitfalls

During the process of this meta-analysis, common issues and pitfalls in selection psychology were encountered. The stakes in ATCO selection are high whether from a state or commercial perspective. Air traffic control can be regalian, military or private and a selection process can be organized by a public administration or by a company. Open Data has not yet reached selection psychology, scientific articles are rare and data is difficult to obtain. Once data is collected, extraction of results can be arduous, the methodology section is generally poor in detail. Nevertheless, the most difficult task is the categorization of tests into predictors. The description of tests is generally very simplistic, the same predictor can be used to design different tests and a number of predictors can be measured with the same test. Composite scores are very difficult to interpret since multiple variables of different nature are averaged together and each article uses different variables. Inaccuracy of description can be used to protect national/commercial secrets. On the other hand, this inaccuracy can be a sign of the greatest problem of psychological measurement. In an ideal world, the theory should guide all aspects of test creation, ranging from item development and selection to score calculation and interpretation (Beaujean & Benson, 2019). In real life, Sijtsma (2012) noted: “items are often constructed guided by best guesses on the basis of whatever theory is available, but also based on intuition [...], tradition [...], and conformity [...]” (p. 790).

Implications for the Research and Practice of Personnel Selection

There are many perspectives for the improvement of ATCO selection. ATCO activities evolve with time but job analysis showed that the same specific cognitive abilities are required for these activities (Eißfeldt, 1991, 2009). Understanding and operationalization of constructs are constantly improving, consequently constructs can be assessed more accurately today than in the past. For example, new methods such physiological measurements are used to measure resilience stress in ATCOs (Ćosić, Popović, et al., 2019; Ćosić, Šarlija, et al., 2019). Some constructs widely used in other professions could be transferred to ATCO selection such as grit in military contexts (Duckworth, 2016; Schimschal, Visentin, Kornhaber, & Cleary, 2020), proficiency or neural efficiency in piloting (Matton, Paubel, & Puma, 2020), (cognitive) stress resistance for pilots (Grassmann, Vlemincx, von Leupoldt, & Van den Bergh, 2016), ability emotional intelligence (O’Connor, Hill, Kaya, & Martin, 2019) and social competence in pilot contexts (Hoermann & Goerke, 2014), vigilance (Robison & Brewer, 2019), cognitive flexibility in CCTV centers (Marois, Hodgetts, Chamberland, Williot, & Tremblay, 2021). In the future, automation (Ferreira & Cañas, 2019) or even adaptive automation (Di Flumeri et al., 2019) is likely to be widespread and pervasive. Trust in automation (Körber, 2018) and technology-acceptance factors (Mudgal, Sharma, Chaturvedi, & Sharma, 2020) could be of interest in ATCO selection. Inside Cattell-Horn-Carroll theory, learning efficiency and retrieval fluency could be of interest in order to select applicants who will learn and strategically retrieve verbal and nonverbal information or ideas stored in long-term memory.

Another target of improvement could be the criterion (Austin & Villanova, 1992). Almost all predictive validity studies use pass-fail criterion and this was the most common criterion used for this meta-analysis. For Broach (2014), the most relevant criterion should be the degree of proficiency in real-duty flight performance for pilots. By analogy, it should be the degree of proficiency in real-duty air traffic control performance for ATCOs. However, a definition of excellent or poor ATCO performance has not been agreed. For example, Eurocontrol has developed a standard criterion for the validation of FEAST tests: the Behavioral Observation Scale (Eißfeldt, 2003) but not all European ANSP use it. Moreover, rating may be affected by structural and cognitive biases affecting the evaluator (Evans & Robinson, 2020). Timing of attrition can be valuable: differences between early attrition and later attrition have been highlighted (Ryan, Sacco, McFarland, & Kriska, 2000). In a highly stressful job (special operation forces), one third of total attrition occurs during the first week. To our knowledge, no information has been published on this subject for ATCOs.

At the psychometric level, development of tests should be guided from theory (Beaujean & Benson, 2019) and more information about such theory or frameworks should be given. A validation study should be performed before introducing a test into selection process and publications about validation studies should include psychometric data such as reliability estimates and restriction-of-range. In ATCO selection, correlation and linear/logistic regression are the gold standard predictivity analyses. Mostly explored in pilot selection, others methods exist such as decision trees and random forests (Giddings, 2020), latent growth mixture (Gomes & Dias, 2015), neural networks (Maroco & Bártolo-Ribeiro, 2013), SVM (Sun et al., 2019). Some papers have been published on the subject with ATCOs (Borack, 1995) but with the democratization and improvement of machine learning, a new interest could emerge for these algorithms. Inspite of this, a simple method can be better than complex ones (Nusinovici et al., 2020) and explicability should be developed: a "black-box" is difficult to certify (Tippins, Oswald, & McPhail, 2021). Based on Pareto-optimality concept (De Corte, Lievens, & Sackett, 2007), elaboration of composite scores and their weightings may be improved in order to consider validity, adverse impact and constraints (De Corte, Sackett, & Lievens, 2010). As used by Lopez, Caffò, Tinella, Postma, and Bosco (2020) for visuo-spatial processing testing, Differential Item Functioning Analysis could help to control adverse impact.

References

References marked with an "a" indicate studies included in the cognitive meta-analysis and a "b" indicate studies included in the non-cognitive meta-analysis.

[a,b] Ackerman, P. L., & Kanfer, R. (1993a). *Development of an ability-based selection instrument for the MARC-Air traffic controller training program*. University of Minnesota, Minneapolis. Retrieved from University of Minnesota, Minneapolis website: https://hdl.handle.net/11299/156850

[a,b] Ackerman, P. L., & Kanfer, R. (1993b). Integrating laboratory and field study for improving selection: Development of a battery for predicting air traffic controller success. *Journal of Applied Psychology*, *78*(3), 413. https://doi.org/ccv5vt

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/ggxvw8

ALMamari, K., & Traynor, A. (2020). Predictive Validity of the Air Force Officer Qualifying Test (AFOQT) for Pilot Performance: A Meta-Analytic Investigation at the Subtest Level. *Aviation Psychology and Applied Human Factors*, 1–12. https://doi.org/gg8zfh

Aust, F., & Barth, M. (2020). *Papaja : Create APA manuscripts with R Markdown*.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology*, *77*(6), 836. https://doi.org/fksqpb

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

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*(1-2), 9–30. https://doi.org/10.1111/1468-2389.00160

Beaujean, A. A., & Benson, N. F. (2019). Theoretically-consistent cognitive ability test development and score interpretation. *Contemporary School Psychology*, *23*(2), 126–137. https://doi.org/gjts39

Beier, M., Kell, H., & Lang, J. (2019). Commenting on the "Great Debate": General Abilities, Specific Abilities, and the Tools of the Trade. *Journal of Intelligence*, *7*(1), 5. https://doi.org/gg8zfw

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/gfpkdx

Binning, J. E., & Barrett, G. V. (1989). Validity of Personnel Decisions: A Conceptual Analysis of the Inferential and Evidential Bases. *American Psychological Association*, *74*(3), 478–494. https://doi.org/b383m5

[a] Bleckley, M. K., Crutchfield, J., King, R. E., Manning, C. A., & Carretta, T. R. (2009). Selection for Aviation Related Careers : Air Traffic Control in the Air Force and the FAA. *2009 International Symposium on Aviation Psychology*, 69–75. Retrieved from https://corescholar.libraries.wright.edu/isap_2009/103

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The Usefulness of Unit Weights in Creating Composite Scores: A Literature Review, Application to Content Validity, and Meta-Analysis. *Organizational Research Methods*, *10*(4), 689–709. https://doi.org/dtx38t

[a,b] Boer, L. C., Harsveld, M., & Hermans, P. H. (1997). The Selective-Listening Task as a Test for Pilots and Air Traffic Controllers. *Military Psychology*, *9*(2), 137–149. https://doi.org/dcr8mk

[a] Boone, J. O. (1980). Toward the development of a new aptitude selection test battery for air traffic control specialists. *Aviation, Space, and Environmental Medicine*, 694–699.

Borack, J. I. (1995). Alternative techniques for predicting success in air controller school. *Military Psychology*, *7*(3), 207–219. https://doi.org/fhjct5

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/gf4nkz

Breaugh, J. A. (2009). The use of biodata for employee selection: Past research and future directions. *Human Resource Management Review*, *19*(3), 219–231. https://doi.org/bh23gn

[a] Broach, D. (1998). *Recovery of the FAA Air Traffic Control Specialist Workforce, 1981-1992.* FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA355135

Broach, D. (2013). *Selection of the next generation of air traffic control specialists: Aptitude requirements for the air traffic control tower cab in 2018.* FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA579808

Broach, D. (2014). Commentary on the article by Weissmuller and Damos: Statistical models and the structure of selection processes in pilot selection systems. *The International Journal of Aviation Psychology*, *24*(2), 119–122. https://doi.org/ggxv4m

Broach, D. (2017). Attrition in US Air Traffic Control Specialist (ATCS) Training: A Review of 50 Years of Data. *19th International Symposium on Aviation Psychology*, 542. Retrieved from https://corescholar.libraries.wright.edu/isap_2017/88

[a] Broach, D. (2019). Innovative ATCS Aptitude Tests : Validity & Utility of the Radar Vectoring Aptitude Test (RVAT). *20th International Symposium on Aviation Psychology.*

[a] Broach, D., Byrne, C. L., & Manning, C. A. (2013). *The Validity of the Air Traffic Selection and Training (AT-SAT) Test Battery in Operational Use* (No.

DOT/FAA/AM-13/3). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA583743

[a] Broach, D., & Manning, C. A. (1994). *Validity of the air traffic control specialist Nonradar Screen as a predictor of performance in radar-based air traffic control training* (No. DOT/FAA/AM-94/09). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA279754

Broach, D., & Manning, C. A. (1997). *Review of Air Traffic Controller Selection: An International Perspective.* (No. DOT/FAA/AM-97/15). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA328993

Brooker, P. (2008). The Überlingen accident: Macro-Level safety lessons. *Safety Science*, *46*(10), 1483–1508. https://doi.org/btk54p

Buckley, J., Seery, N., & Canty, D. (2018). A heuristic framework of spatial ability: A review and synthesis of spatial factor literature to support its translation into STEM education. *Educational Psychology Review*, *30*(3), 947–972. https://doi.org/gdmxvh

Bühner, M., König, C. J., Pick, M., & Krumm, S. (2006). Working memory dimensions as differential predictors of the speed and error aspect of multitasking performance. *Human Performance*, *19*(3), 253–275. https://doi.org/bc3sdd

Caemmerer, J. M., Keith, T. Z., & Reynolds, M. R. (2020). Beyond individual intelligence tests: Application of Cattell-Horn-Carroll theory. *Intelligence*, *79*, 101433. https://doi.org/ggtgrr

[a] Carretta, Thomas R., & King, R. E. (2008). Improved military air traffic controller selection methods as measured by subsequent training performance. *Aviation Space and Environmental Medicine*, *79*(1), 36–43. https://doi.org/b6bsfv

[a] Carretta, Thomas R., & Siem, F. M. (1999). Determinants of enlisted air traffic controller success. *Aviation, Space, and Environmental Medicine*, *70*(9), 910–918.

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

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1. https://doi.org/fs6ptd

[a,b] Chappelle, W., Thompson, W., Goodman, T., Bryan, C. J., & Reardon, L. (2015). The Utility of Testing Noncognitive Aptitudes as Additional Predictors of Graduation From US Air Force Air Traffic Controller Training. *Aviation Psychology and Applied Human Factors*, *5*(2), 93–103. https://doi.org/ggvk22

[a] Chiles, W. D., & West, G. (1974). *Multiple task performance as a predictor of the potential of air traffic controller trainees: A follow-up study.* (No. DOT/FAA/AM-74/10). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://rosap.ntl.bts.gov/view/dot/21056

Christiansen, N. D., Robie, C., Burns, G. N., Loy, R. W., Speer, A. B., & Jacobs, R. R. (2021). Effects of applicant response distortion on the relationship between personality trait scores and cognitive ability. *Personality and Individual Differences*, *171*, 110542. https://doi.org/ghzvsb

[a,b] Cobb, B. B. (1964). Problems in Air Traffic Management: Identification and potential of aptitude test measures for selection of tower air traffic controller trainees. V. *Aerospace Medicine and Human Performance*, *35*(11), 1019–1027.

[a] Cobb, B. B. (1968). *A Comparative Study of Air Traffic Trainee Aptitude-Test Measures Involving Navy, Marine Corps and FAA Controllers* (No. DOT/FAA/AM-

68/14). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/1960s/media/AM68-14.pdf

[a] Cobb, B. B. (1971). *Air traffic aptitude test measures of military and FAA controller trainees.* (No. DOT/FAA/AM-71/40). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://rosap.ntl.bts.gov/view/dot/20977

[b] Cobb, B. B., Young, C. L., & Rizzuti, B. L. (1976). *Education as a factor in the selection of air traffic controller trainees* (No. DOT/FAA/AM-76/6). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA031880

[a] Cobb, B., & Mathews, J. (1972). *A Proposed New Test for Aptitude Screening of Air Traffic Controller Applicants. FAA Office of Aviation Medicine Report No* (No. DOT/FAA/AM-72/18). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/1970s/media/AM72-18.pdf

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences.* Academic press.

Cokely, D. E. T. (2018). Decision Making Skill: From Intelligence to Numeracy and Expertise. In *Cambridge handbook of expertise and expert performance* (pp. 476–505).

[a,b] Collins, W. E., Nye, L. G., & Manning, C. A. (1992). Poststrike air traffic control trainees: Biodemographic predictors of success in selection and screening. *The International Journal of Aviation Psychology*, *2*(3), 213–223. https://doi.org/fcd37b

[a,b] Conzelmann, K., & Keye, D. (2014). Which Aspects of a Semistructured Interview, Besides Cognitive Ability Tests, Contribute Incrementally to Predicting the

Training Success of Air Traffic Controller Trainees? *International Journal of Selection and Assessment*, *22*(3), 240–252. https://doi.org/gf5xcb

Corr, P. J., & Matthews, G. (2020). *The Cambridge handbook of personality psychology*. Cambridge University Press.

Ćosić, K., Popović, S., Šarlija, M., Mijić, I., Kokot, M., Kesedžić, I., … Zhang, Q. (2019). New tools and methods in selection of air traffic controllers based on multimodal psychophysiological measurements. *Ieee Access*, *7*, 174873–174888. https://doi.org/ggvk7b

Ćosić, K., Šarlija, M., Ivkovic, V., Zhang, Q., Strangman, G., & Popović, S. (2019). Stress resilience assessment based on physiological features in selection of air traffic controllers. *IEEE Access*, *7*, 41989–42005. https://doi.org/gghjm4

Costa, P. T., & McCrae, R. R. (1999). A five-factor theory of personality. *Handbook of Personality, 2nd Edn. Guilford Press, New York*, 139–153.

Dahlke, J. A., & Wiernik, B. M. (2019). Psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*, *43*(5), 415–416. https://doi.org/gfgt9t

Dahlke, J. A., & Wiernik, B. M. (2020). Not restricted to selection research: Accounting for indirect range restriction in organizational research. *Organizational Research Methods*, *23*(4), 717–749. https://doi.org/gf9crt

[a] Dailey, J. T., & Pickrel, E. W. (1977). *Development of new selection tests for air traffic controllers* (No. DOT/FAA/AM-77/25). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA049049

[a] Damitz, M., & Eisfeldt, H. (2004). The Relevance of General Cognitive Ability (g) for Training Success of Controllers. In *Aviation Psychology: Practice and Research* (p. 135). Routledge.

Dang, C.-P., Braeken, J., Colom, R., Ferrer, E., & Liu, C. (2015). Do processing speed and short-term storage exhaust the relation between working memory capacity and intelligence? *Personality and Individual Differences*, *74*, 241–247. https://doi.org/f6xc38

Darr, W. (2011). Military Personality Research: A Meta-Analysis of the Self Description Inventory. *Military Psychology*, *23*(3), 272–296. https://doi.org/ch7pn6

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, *92*(5), 1380. https://doi.org/dnxwd5

De Corte, W., Sackett, P., & Lievens, F. (2010). Selecting Predictor Subsets: Considering validity and adverse impact: Selecting Predictor Subsets. *International Journal of Selection and Assessment*, *18*(3), 260–270. https://doi.org/bpqtqz

[a,b] Dean, M., & Broach, D. (2012). *Development, Validation, and Fairness of a Biographical Data Questionnaire for the Air Traffic Control Specialist (ATCS) Occupation*. FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://rosap.ntl.bts.gov/view/dot/25441

[a] Della Rocco, P. S., Manning, C. A., & Wing, H. (1990). *Selection of air traffic controllers for automated systems: Applications from current research* (No. DOT/FAA/AM-90/13). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA230058

[a] Della Rocco, P. S., Milburn, N., & Mertens, H. W. (1992). *Comparison of performance on the Shipley Institute of Living Scale, air traffic control specialist selection test, and FAA Academy screen* (No. DOT/FAA/AM-92/30). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA259594

Di Flumeri, G., De Crescenzio, F., Berberian, B., Ohneiser, O., Kramer, J., Aricò, P., … Piastra, S. (2019). Brain–computer interface-based adaptive automation to prevent out-of-the-loop phenomenon in air traffic controllers dealing with highly automated systems. *Frontiers in Human Neuroscience*, *13*, 296. https://doi.org/10.3389/fnhum.2019.00296

Duckworth, A. L. (2016). *Grit: The power of passion and perseverance* (Vol. 234). Scribner New York, NY.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*(19), 7716–7720. https://doi.org/bmd

[b] Dwan, A. (2011). *The Effects of Age, Sex and Education Level on Air Traffic Control Training Outcomes* (PhD thesis). University of Canterbury. Psychology.

Eißfeldt, H. (1991). Automation in ATC: How Does It Affect the Selection of Controllers? In *Automation and Systems Issues in Air Traffic Control* (pp. 461–465). Springer.

Eißfeldt, H. (2003). The development and validation of a common FEAST criterion. *Proceedings of the Second EUROCONTROL Selection Seminar*.

Eißfeldt, H. (2009). Aviator 2030-Ability requirements in future ATM Systems. *2009 International Symposium on Aviation Psychology*, 112.

Eurocontrol. (2020). *First European Air traffic controller Selection Test*.

Evans, L., & Robinson, G. L. (2020). Evaluating Our Evaluations: Recognizing and Countering Performance Evaluation Pitfalls. *Military Review*, *100*(1), 89. Retrieved from https://digitalcommons.usmalibrary.org/usma_research_papers/316/

FAA. (2021). *FAA Aerospace Forecast 2021-2041*. Retrieved from https://www.faa.gov/data_research/aviation/aerospace_forecasts/media/Forecast_Highlights.pdf

Falch, T., & Sandgren Massih, S. (2011). The effect of education on cognitive ability. *Economic Inquiry*, *49*(3), 838–856. https://doi.org/10.1111/j.1465-7295.2010.00

312.x

Farmer, W. L. (2002). *Characteristics of biodata keys as a function of scaling method, sample size, and criterion.* (PhD thesis, University of Oklahoma). University of Oklahoma. Retrieved from https://hdl.handle.net/11244/451

[a] FEAST. (2010). *Validation of the First European Air Traffic Controller Selection Tests (FEAST)* (pp. 1–110). Brussels, Belgium: Eurocontrol.

[a] FEAST. (2016). *Validation of the Dynamic ATC Radar Test.* Brussels, Belgium: Eurocontrol.

[a] FEAST. (2019). *Validation of the FEAST MULTI-PASS Test.* Brussels, Belgium: Eurocontrol.

Federal Aviation Administration. (2020). *The Air Traffic Controller Workforce 2020 / 2029.* Retrieved from https://www.faa.gov/air_traffic/publications/controller_staffing/

Ferreira, P. N., & Cañas, J. J. (2019). Assessing operational impacts of automation using functional resonance analysis method. *Cognition, Technology & Work*, *21*(3), 535–552. https://doi.org/ggkqht

Fiori, M., & Vesely-Maillefer, A. K. (2018). Emotional intelligence as an ability: Theory, challenges, and new directions. In *Emotional intelligence in education* (pp. 23–47). Springer.

Flanagan, D., & McDonough, E. (2018). *Contemporary intellectual assessment: Theories, tests, and issues.* Guilford Publications.

[a,b] Fox, K. D. (2011). *Prediction of Air Traffic Controller Trainee Selection and Training Success Using Cognitive Ability and Biodata* (PhD thesis, Walden University). Walden University. Retrieved from https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=2119&context=dissertations

Frischkorn, G. T., Schubert, A.-L., & Hagemann, D. (2019). Processing speed, working memory, and executive functions: Independent or inter-related predictors of general intelligence. *Intelligence*, *75*, 95–110. https://doi.org/gf3sxs

Ganuthula, V. R. R., & Sinha, S. (2019). The Looking Glass for Intelligence Quotient Tests: The Interplay of Motivation, Cognitive Functioning, and Affect. *Frontiers in Psychology*, *10*, 2857. https://doi.org/ggxvvz

Giddings, A. C. (2020). *Predicting Pilot Success Using Machine Learning* (PhD thesis, Air Force Institute of Technology). Air Force Institute of Technology. Retrieved from https://scholar.afit.edu/etd/3196

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/f84bhv

Gomes, A., & Dias, J. G. (2015). Improving the selection of air force pilot candidates using latent trajectories: An application of latent growth mixture modeling. *The International Journal of Aviation Psychology*, *25*(2), 108–121. https://doi.org/ggxv46

Grassmann, M., Vlemincx, E., von Leupoldt, A., & Van den Bergh, O. (2016). The role of respiratory measures to assess mental load in pilot selection. *Ergonomics*, *59*(6), 745–753. https://doi.org/ggxvv2

Gutman, L. M., & Schoon, I. (2013). The impact of non-cognitive skills on outcomes for young people. *Education Endowment Foundation*, *59*(22.2), 2019. Retrieved from http://hdl.voced.edu.au/10707/287500

Hattrup, K. (2012). Using composite predictors in personnel selection. *The Oxford Handbook of Personnel Assessment and Selection*, 297–319. https://doi.org/ghmjv8

Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical Research Ed.)*, *327*(7414), 557–560. https://doi.org/dhbjj6

Hoermann, H.-J., & Goerke, P. (2014). Assessment of Social Competence for Pilot Selection. *The International Journal of Aviation Psychology*, *24*(1), 6–28. https:

//doi.org/ggxv5b

Hoermann, H.-J., Noser, P., & Stelling, D. (2018). A Psychometrician's Dream: Test Validities in an Unrestricted Sample of Future Swiss Air Force Pilots. *33rd Conference of the European Association for Aviation Psychology (EAAP)*, 6. Retrieved from https://elib.dlr.de/121976/

Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities* (PhD thesis). University of Illinois.

Huck, J. (2020). *Who makes better decisions? The relative importance of numeracy and cognitive abilities for elements of decision making* (PhD thesis, University of Oklahoma). University of Oklahoma. Retrieved from https://hdl.handle.net/11244/325348

Hunter, D. R., & Burke, E. F. (1994). Predicting aircraft pilot-training success: A meta-analysis of published research. *The International Journal of Aviation Psychology*, *4*(4), 297–313. https://doi.org/dmz43h

International Civil Aviation Organization. (2011). *ICAO study reveals strong demand for qualified aviation personnel up to 2030*. Retrieved from https://www.icao.int/Newsroom/Pages/FR/icao-study-reveals-strong-demand-for-qualified-aviation-personnel-up-to-2030_FR.aspx

Jeong, Y. R., Christiansen, N. D., Robie, C., Kung, M.-C., & Kinney, T. B. (2017). Comparing applicants and incumbents: Effects of response distortion on mean scores and validity of personality measures. *International Journal of Selection and Assessment*, *25*(3), 311–315. https://doi.org/ghnwg2

Johnson, J. F., Barron, L. G., Carretta, T. R., & Rose, M. R. (2017). Predictive validity of spatial ability and perceptual speed tests for aviator training. *The International Journal of Aerospace Psychology*, *27*(3-4), 109–120. https://doi.org/ggxv44

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An

Integrative/Aptitude-Treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, *74*(4), 657. https://doi.org/bgq35m

Kanfer, R., & Ackerman, P. L. (2005). Work competence. *Handbook of Competence and Motivation*, 336–353.

[a,b] Kelley, R. (2012). *Relationship between air traffic selection and training battery test scores and composites scores in the initial en route air traffic control qualification training course at the FAA* (PhD thesis, Oklahoma State University). Oklahoma State University. Retrieved from https://eric.ed.gov/?id=ED547900

King, R. E., Manning, C. A., Schroeder, D. J., Carretta, T. R., Rathje, H., & Myhr, R. (2007). A Worldwide Review of Selection for Air Traffic Control Personnel. *2007 International Symposium on Aviation Psychology*, 7. Retrieved from https://corescholar.libraries.wright.edu/isap_2007/79

Kisamore, J. L., & Brannick, M. T. (2008). An Illustration of the Consequences of Meta-Analysis Model Choice. *Organizational Research Methods*, *11*(1), 35–53. https://doi.org/dcvjqb

Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. *Congress of the International Ergonomics Association*, 13–30. Springer. https://doi.org/10.1007/978-3-319-96074-6_2

Koslowsky, M., & Sagie, A. (1993). On the efficacy of credibility intervals as indicators of moderator effects in meta-analytic research. *Journal of Organizational Behavior*, *14*(7), 695–699. https://doi.org/bh4pgz

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151–177. https://doi.org/gd3vr6

Kovacs, K., & Conway, A. R. (2019). A unified cognitive/differential approach to human intelligence: Implications for IQ testing. *Journal of Applied Research in Memory and Cognition*, *8*(3), 255–272. https://doi.org/ggbhnx

Lang, J. W. B., & Kell, H. J. (2020). General mental ability and specific abilities: Their relative importance for extrinsic career success. *Journal of Applied Psychology*, *105*(9), 1047–1061. https://doi.org/ggf2pk

Lievens, F., Sackett, P. R., & Zhang, C. (2020). Personnel selection: A longstanding story of impact at the individual, firm, and societal level. *European Journal of Work and Organizational Psychology*, 1–12. https://doi.org/ghmvj9

Lopez, A., Caffò, A. O., Tinella, L., Postma, A., & Bosco, A. (2020). Studying Individual Differences in Spatial Cognition Through Differential Item Functioning Analysis. *Brain Sciences*, *10*(11), 774. https://doi.org/ghhn82

Lubinski, D. (2010). Spatial ability and STEM: A sleeping giant for talent identification and development. *Personality and Individual Differences*, *49*(4), 344–351. https://doi.org/dk3tdn

Lüdecke, D. (2019). *Esc: Effect size computation for meta analysis (version 0.5.1)*.

Major, J. T., Johnson, W., & Bouchard Jr, T. J. (2011). The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent g. *Intelligence*, *39*(5), 418–433. https://doi.org/dhq7w6

Mammadov, S. (2021). The Big Five Personality Traits and Academic Performance: A Meta-Analysis. *Journal of Personality*. https://doi.org/gk9x9c

[b] Manley, G. G. (2003). *Investigating the linear predictor - criterion assumption of biodata scaling* (PhD thesis, University of Oklahoma). University of Oklahoma. Retrieved from https://hdl.handle.net/11244/684

[a] Manning, C. A., Della Rocco, P. S., & Bryant, K. D. (1989). *Prediction of success in FAA air traffic control field training as a function of selection and screening test performance.* (No. DOT/FAA/AM-89/6). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://rosap.ntl.bts.gov/view/dot/21315

Maroco, J., & Bártolo-Ribeiro, R. (2013). Selection of air force pilot candidates: A case study on the predictive accuracy of discriminant analysis, logistic regression, and four neural network types. *The International Journal of Aviation Psychology*, *23*(2), 130–152. https://doi.org/ggxv35

Marois, A., Hodgetts, H. M., Chamberland, C., Williot, A., & Tremblay, S. (2021). Who can best find Waldo? Exploring individual differences that bolster performance in a security surveillance microworld. *Applied Cognitive Psychology*, acp.3837. https://doi.org/gjxw5n

Martinussen, M., & Hunter, D. R. (2017). *Aviation psychology and human factors*. CRC Press.

Martinussen, M., Jenssen, M., & Joner, A. (2000). Selection of air traffic controllers : Some preliminary findings from a meta-analysis of validation studies. *Proceedings of the 24th EAAP (European Association for Aviation Psychology) Conference*.

Matton, N., Paubel, P.-V., & Puma, S. (2020). Toward the use of pupillary responses for pilot selection. *Human Factors*, 0018720820945163. https://doi.org/gg9p69

Matton, N., Vautier, S., & Raufaste, É. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence*, *37*(4), 412–421. https://doi.org/fdktn8

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In *Emotional development and emotional intelligence: Educational implications.* Basic Books.

McGill, R. J., & Dombrowski, S. C. (2019). Critically reflecting on the origins, evolution, and impact of the Cattell-Horn-Carroll (CHC) model. *Applied Measurement in Education*, *32*(3), 216–231. https://doi.org/gg8zf6

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 136–181). The Guilford Press.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, *6*(7), e1000097. https://doi.org/bq3jpc

Morath, R., Quartetti, D., Bayless, A., & Archambault, C. (2001). Air traffic controller job analysis. *Documentation of Validity for the AT-SAT Computerized Test Battery*, *1*, 7–18.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika*, *8*(3), 161–168. https://doi.org/10.1007/BF02288700

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., … Lukaszewski, A. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits. *European Journal of Personality*, *34*(6), 1175–1201. https://doi.org/10.1002/per.2311

Mudgal, S. K., Sharma, S. K., Chaturvedi, J., & Sharma, A. (2020). Brain computer interface advancement in neurosciences: Applications and issues. *Interdisciplinary Neurosurgery*, *20*, 100694. https://doi.org/ggzbmw

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., … Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/ghhn84

[b] Nye, L. G., Schroeder, D. J., & Dollar, C. S. (1994). *Relationships of type a behavior with biographical characteristics and training performance of air traffic controllers* (No. DOT/FAA/AM-94/13). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://apps.dtic.mil/sti/citations/ADA283813

O'Boyle, E. H., Humphrey, R. H., Pollack, J. M., Hawver, T. H., & Story, P. A. (2011). The relation between emotional intelligence and job performance: A meta-analysis. *Journal of Organizational Behavior*, *32*(5), 788–818. https://doi.org/dng9fc

O'Connor, P. J., Hill, A., Kaya, M., & Martin, B. (2019). The Measurement of

Emotional Intelligence: A Critical Review of the Literature and Recommendations for Researchers and Practitioners. *Frontiers in Psychology*, *10*, 1116. https://doi.org/gf3sdn

[a,b] Oakes, D. W., Ferris, G. R., Martocchio, J. J., Buckley, M. R., & Broach, D. (2001). Cognitive ability and personality predictors of training program skill acquisition and job performance. *Journal of Business and Psychology*, *15*(4), 523–548. https://doi.org/fp2cjm

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Brennan, S. E. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Bmj*, *372*. https://doi.org/10.1136/bmj.n71

Pavel, S. (2012). The Evolution of the ATC Selection and Training Process. *Journal of Aviation/Aerospace Education & Research*, *21*(2). https://doi.org/ggzjkj

Pecena, Y., Gayraud, K., & Eißfeldt, H. (2019). Aiming High: Maximizing Predictive Validity of Selection Processes for Air Traffic Controller Candidates. In *Pilot Selection: Psychological Principles and Practice*. CRC Press.

[a,b] Pecena, Y., Keye, D., Conzelmann, K., Grasshoff, D., Maschke, P., Heintz, A., & Eisfeldt, H. (2013). Predictive Validity of a Selection Procedure for Air Traffic Controller Trainees. *Aviation Psychology and Applied Human Factors*, *3*(1), 19–27. https://doi.org/ggvk24

Petrides, K., & Furnham, A. (2000). On the dimensional structure of emotional intelligence. *Personality and Individual Differences*, *29*(2), 313–320. https://doi.org/d7c8vt

Petrides, K., Pita, R., & Kokkinaki, F. (2007). The location of trait emotional intelligence in personality factor space. *British Journal of Psychology*, *98*(2), 273–289. https://doi.org/ddzfcs

[a] Pierce, L., & Bleckley, M. K. (2013). The Utility of AT-SAT in Hiring Graduates

of an Air-Traffic Collegiate Training Initiative Program. *17th International Symposium on Aviation Psychology*, 663–668. Retrieved from https://corescholar.libraries.wright.edu/isap_2013/4

[b] Pierce, L., Broach, D., Bleckley, M. K., & Byrne, C. L. (2013). Using biodata to select air traffic controllers paper. *Proceedings of the Human Factors and Ergonomics Society*, 823–827. https://doi.org/ggvk26

[a,b] Pierce, L., Broach, D., Bleckley, M. K., & Byrne, C. L. (2014). Using Biodata to Select Air Traffic Controllers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*, 823–827. https://doi.org/ggvk26

[a,b] Pierce, L., Byrne, C., Beben, M., & Pfleiderer, E. (2016). Using a Multidimensional Pairwise Preference Measure of Personality to Predict Training Success of Air Traffic Control Specialists. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*, 1314–1318. https://doi.org/ggvk23

Ployhart, R. E., Schneider, B., & Schmitt, N. (2005). *Staffing organizations: Contemporary practice and theory*. CRC Press.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria.

Redding, R. E., Cannon, J. R., & Seamster, T. L. (1992). Expertise in Air Traffic Control (ATC): What is it, and how Can we Train for It? *Proceedings of the Human Factors Society Annual Meeting*, *36*(17), 1326–1330. https://doi.org/c3bvzd

Robison, M. K., & Brewer, G. A. (2019). *Individual differences in vigilance: Implications for measuring sustained attention and its association with other cognitive abilities and psychological constructs* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/nc9b6

[a] Rock, D., Dailey, J., Ozur, H., Boone, J., & Pickrel, E. (1984). Validity and utility of the ATC experimental tests battery. Study of Academy trainees, 1982. In

Selection of air traffic controllers. FAA Office of Aerospace Medicine. Retrieved from https://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/1980s/media/AM84-02.pdf

Roe, R. A. (2005). The design of selection systems: Context, principles, issues. In *The Blackwell Handbook of Personnel Selection*. Blackwell Publishing.

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity : Updating and integrating some classic litterature. *Personnel Psychology*, *58*(4), 1009–1037. https://doi.org/fq8jzm

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, 1–7. https://doi.org/b2pp6s

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, S. D. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology*, *85*(2), 163. https://doi.org/bz33tz

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000994

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition and Personality*, *9*(3), 185–211. https://doi.org/d7s8wh

Schimschal, S. E., Visentin, D., Kornhaber, R., & Cleary, M. (2020). Grit: A concept analysis. *Issues in Mental Health Nursing*, 1–11. https://doi.org/10.1080/01612840.2020.1814913

Schmank, C. J., Goring, S. A., Kovacs, K., & Conway, A. R. (2021). Investigating the structure of intelligence using latent variable and psychometric network modeling: A commentary and reanalysis. *Journal of Intelligence*, *9*(1), 8. https://doi.org/gh28p5

Schmidt, F. (2017). Statistical and measurement pitfalls in the use of meta-regression in meta-analysis. *Career Development International*, *22*(5), 469–476. https://doi.org/gcnjf4

Schmidt, F., & Hunter, J. E. (1998). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings. *Psychological Bulletin*, *124*(2), 13. https://doi.org/fd8vxp

Schmidt, F., & Hunter, J. E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (Third Edition). 55 City Road, London.

Schneider, W., & McGrew, K. (2012). The Cattell-Horn-Carroll model of intelligence. In *Contemporary intellectual assessment: Theories, tests, and issues* (p. 119). The Guilford Press.

Schneider, W., & McGrew, K. (2018). The Cattell Horn Carroll Theory of Cognitive Abilities. In *Contemporary Intellectual Assessment Fourth Edition*.

Schröder, V. S., Heimann, A. L., Ingold, P. V., & Kleinmann, M. (2021). Enhancing Personality Assessment in the Selection Context: A Study Protocol on Alternative Measures and an Extended Bandwidth of Criteria. *Frontiers in Psychology*, *12*, 643690. https://doi.org/gj3wfp

[a,b] Schroeder, D. J., Broach, D., & Young, W. C. (1993). *Contribution of personality to the prediction of success in initial air traffic control specialist training* (No. DOT/FAA/AM-93/4). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://rosap.ntl.bts.gov/view/dot/21349

[a] Schroeder, D. J., Dollar, C. S., & Nye, L. G. (1990). *Correlates of two experimental tests with performance in the FAA Academy Air Traffic Control Nonradar Screen Program* (No. DOT/FAA/AM-90/8). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://rosap.ntl.bts.gov/view/dot/21310

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *The-*

ory & Psychology, *22*(6), 786–809. https://doi.org/10.1177/0959354312454353

Skyguide. (2021). *Become a certified Air Traffic Controller.* Retrieved from https://www.skyguide.ch/en/work-at-skyguide/become-an-atco/

Song, G., Wu, J., & Wang, S. (2019). Recent Ten Years of Biodata Research: Hotspots, Progress and Prospects. *Proceedings of the 4th International Conference on Economy, Judicature, Administration and Humanitarian Projects (JAHP 2019).* Kaifeng, China: Atlantis Press. https://doi.org/gjpxkt

Speer, A. B., Tenbrink, A. P., Wegmeyer, L. J., Sendra, C. C., Shihadeh, M., & Kaur, S. (2021). Meta-analysis of biodata in employment settings: Providing clarity to criterion and construct-related validity estimates. *Journal of Applied Psychology.* https://doi.org/10.1037/apl0000964

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., … Ullmann, T. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680–17687. https://doi.org/gg86vz

[a] Stoker, P., Hunter, D. R., Batchelor, C. L., & Curran, L. T. (1987). *Air Traffic Controller Trainee Selection.* (No. AFHRL/TP/87/19). Air Force Human Resources Lab Brooks AFB TX. Retrieved from Air Force Human Resources Lab Brooks AFB TX website: https://apps.dtic.mil/sti/citations/ADA187497

Sun, J., Xiao, X., Cheng, S., Shen, C., Ma, J., & Hu, W. (2019). Study on Pilot Personality Selection with an SVM-Based Classifier. In S. Long & B. S. Dhillon (Eds.), *Man-Machine-Environment System Engineering* (Vol. 527, pp. 3–10). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-2481-9_1

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). *Scientific, Legal, and Ethical Concerns about AI-Based Personnel Selection Tools: A Call to Action* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/6gczw

[a] Trites, D. K. (1961). *Problems in air traffic management. I., Longitudinal prediction*

of effectiveness of air traffic controllers. (No. DOT/FAA/AM-61/1). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website: https://rosap.ntl.bts.gov/view/dot/20679

[a,b] Trites, D. K., & Cobb, B. B. (1963). *Problems in Air Traffic Management IV: Comparison of Pre-Employment, Job-Related Experience with Aptitude Tests as Predictors of Training and Job Performance of Air Traffic Control Specialists* (No. DOT/FAA/AM-63/31). FAA Office of Aerospace Medicine.

U.S. Department of Transportation. (2020). *Audit Initiated of FAA's Air Traffic Controller Hiring Reform.* Retrieved from https://www.oig.dot.gov/library-item/38011

[b] VanDeventer, A. D. (1983). Biographical profiles of successful and unsuccessful air traffic control specialist trainees. In *Three studies of biographical factors associated with success in air traffic control specialist Screening/Training at the FAA Academy*: *Vol. DOT/FAA/AM-83/6.* FAA Office of Aerospace Medicine. Retrieved from https://rosap.ntl.bts.gov/view/dot/21231

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/gckfpj

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/fs53xp

Viswesvaran, C., Ones, D. S., Schmidt, F., Le, H., & Oh, I.-S. (2014). Measurement Error Obfuscates Scientific Knowledge: Path to Cumulative Knowledge Requires Corrections for Unreliability and Psychometric Meta-Analyses: Measurement error and scientific knowledge. *Industrial and Organizational Psychology*, *7*(4), 507–518. https://doi.org/5k5

[a,b] Walker, K. M., Brou, R., & Farmer, W. L. (2014). *Suitability Screening Test for Marine Corps Air Traffic Controllers Phase III: Non-Cognitive Test Validation*

and Cognitive Test Prototype (No. NPRST/TR/14/3). Navy Personnel Research, Studies, and Technology. Retrieved from Navy Personnel Research, Studies, and Technology website: https://apps.dtic.mil/sti/citations/ADA602372

[a,b] Walker, K. M., Farmer, W. L., & Roberts, R. C. (2013). *Suitability Screening Test for Marine Corps Air Traffic Controllers Phase II* (No. NPRST/TR/13/2). Navy Personnel Research, Studies, and Technology. Retrieved from Navy Personnel Research, Studies, and Technology website: https://apps.dtic.mil/sti/citations/ADA578108

[a,b] Walker, K., Farmer, W., & Roberts, R. (2012). *Suitability Screening Test for Marine Corps Air Traffic Controllers* (No. NPRST/TR/13/1). Navy Personnel Research, Studies, and Technology. Retrieved from Navy Personnel Research, Studies, and Technology website: https://apps.dtic.mil/sti/citations/ADA574499

Wasserman, J. D. (2019). Deconstructing CHC. *Applied Measurement in Education*, *32*(3), 249–268. https://doi.org/gg8zgn

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94–123. https://doi.org/ggj6xq

Wiernik, B. M., Ones, D. S., Marlin, B. M., Giordano, C., Dilchert, S., Mercado, B. K., … Ellis, B. (2020). Using mobile sensors to study personality dynamics. *European Journal of Psychological Assessment*. https://doi.org/ghzvrw

Wood, J. 'Andy'. (2008). Methodology for Dealing With Duplicate Study Effects in a Meta-Analysis. *Organizational Research Methods*, *11*(1), 79–95. https://doi.org/bgjgs3

[a] Young, W. C., Broach, D., & Farmer, W. L. (1996). *Differential Prediction of FAA Academy Performance on the Basis of Gender and Written Air Traffic Control Specialist Aptitude Test Scores.* (No. DOT/FAA/AM-96/13). FAA Office of Aerospace Medicine. Retrieved from FAA Office of Aerospace Medicine website:

https://apps.dtic.mil/sti/citations/ADA308354

Zhang, C., & Kuncel, N. R. (2020). Moving Beyond the Brag Sheet: A Meta-Analysis of Biodata Measures Predicting Student Outcomes. *Educational Measurement: Issues and Practice*, *39*(3), 106–121. https://doi.org/ghk2qh