Selection in Aviation

A European Association for Aviation Psychology Report

This report was prepared by:

J. Eaglestone, D. Damos, H.-J. Hörmann, K. Stadler, J. Wium

As members of the EAAP Working Group on Selection in Aviation

EAAP

EUROPEAN ASSOCIATION FOR AVIATION PSYCHOLOGY

Preface

At the beginning of 2019, at a board meeting in Vienna, the EAAP Board voted to endorse working groups on several important topics. These were subjects that were deemed of interest to EAAP members through the numerous questions, comments and requests received by the Board. One of the topics EAAP often receives questions on is the selection of aviation personnel. Based on this a working group was formed with EAAP member volunteers. The working group itself decided the way forward.

In September 2019 the EAAP Working Group on Selection in Aviation met for the first time, also in Vienna. They decided that a white paper and a survey on current practice would be the best way to help members. The working group continued to meet online for the next 2.5 years to produce this report. Although the working group originally envisioned producing a white paper, the amount of material available on selection resulted in a much larger document. The EAAP Working Group on Selection in Aviation was led by Jennifer Eaglestone (NLR). Members (in alphabetical order) were Diane Damos (Damos Aviation Services, Inc.), Hans-Jürgen Hörmann (Independent, retired DLR), Karien Stadler (Independent) and Johann Wium (Independent).

I would like to thank all of those involved in producing this truly valuable document.

Gunnar Steinhardt,

EAAP President

Authors

Dr. Diane L. Damos (Independent, Damos Aviation Services, Inc.)

Dr. Damos has been involved in pilot selection for over 45 years. She received her doctorate from the University of Illinois in aviation psychology. She was a professor of human factors for 19 years, most recently at the Institute of Safety and Systems Management at the University of Southern California. In 1995 she founded Damos Aviation Services, Inc., a company specializing in pilot selection.

Dr. Damos has conducted research on pilot selection for both the US Air Force and the Navy. She has been a consultant for air carriers, training schools, and governments in the US, Africa, the West Indies, Canada, and Asia. She has taught courses on pilot selection in Taiwan, South Africa, Spain, and Canada as well as in the US.

Jennifer Eaglestone (Senior Consultant, Netherlands Aerospace Centre)

Jennifer is an EAAP (European Association for Aviation Psychology) Board member, accredited Aviation Psychologist, and a senior consultant with the Royal Netherlands Aerospace Centre (NLR). Prior to joining NLR, she worked for many years as a selection expert for high-stress positions firstly with the Netherlands Ministry of Defence and later at Air Traffic Control The Netherlands (LVNL). Jennifer holds a Master's degree in Social and Organizational Psychology from Utrecht University.

Dr. Hans-Jürgen Hörmann (Independent, ret. German Aerospace Center, DLR)

Hans-Jürgen received his Ph.D. in Applied Psychology from the Free University in Berlin and has accumulated over 35 years of research experience in aviation human factors. As a Senior Human Factors Scientist, he was involved in pilot and air-traffic controller selection for the German Aerospace Center in Europe, the Middle East and the Far East. As a Technical Fellow for Safety and Human Factors, he spent three years at Boeing's Research & Technology Center in Madrid and he published over two hundred research papers. His most recent papers are on levels of automation, fatigue and alcohol on human performance in the cockpit. He was awarded a Fellowship of the Royal Aeronautical Society and has been an EAAP accredited Aviation Psychologist since 1996. Hans-Jürgen currently serves as Domain Editor for the International Journal of Aerospace Psychology.

Karien Stadler (Independent and co-founder of AeroAssess FZCO)

Karien started her career in the military and occupied a range of leadership roles, focusing on the selection and development of military leaders and elite soldiers. In 2004, having reached the rank of Lieutenant Colonel, she made a career transition to the corporate world. She worked as HR professional and Industrial Psychologist in South Africa, Saudi Arabia, Sudan and the United Arab Emirates which enabled her to engage with clients across 85 countries, delivering global enterprise talent identification and development programmes.

She recently co-founded AeroAssess - an aviation consultancy business after a tenure of five years as senior psychologist at Emirates Airlines. Karien specialises in pilot selection as well as other talent management interventions within the aviation industry. She holds a Master's degree, is a licensed Industrial Psychologist with the Health Professions Council of South Africa and an EAAP accredited Aviation Psychologist.

Jóhann Wium (Consultant)

Jóhann Wium is an EAAP accredited Aviation Psychologist, a chartered Occupational Psychologist with the BPS (British Psychological Society) and has a Master's degree in Industrial/Organizational Psychology from NYU (New York University). From 2013 to 2022, Jóhann worked for Isavia, Iceland's airport operator and air navigation service provider. While there he worked in a number of different roles (HR consultant, HR and Security Manager, Aviation Psychologist and more) and over-saw the selection of student Air Traffic Controllers for both Iceland and Greenland as well as multiple other operational roles.

Copyright & Disclaimer

Table of Contents

1. Introduction

Jennifer Eaglestone

The EAAP report on Selection in Aviation was written by EAAP members for EAAP members with an interest in selection in aviation. Although some basic knowledge of psychology will help in reading this report, it has been written to be as accessible as possible for anyone interested in selection in aviation. The authors and EAAP would however like to emphasize the fact that having this report in hand does not make you a qualified selection practitioner. It is meant as a point of reference for those interested in the topic, providing an overview of the various aspects involved. A selection procedure should always be designed and overseen by qualified and experienced personnel with an understanding of aviation psychology.

This report takes you from reasons for selection, to the history of selection in aviation, to recommendations for designing, validating and implementing your own selection procedure. The report also includes two appendices, the first is a checklist to assist selection specialists with the technical aspects of a selection procedure; the second presents the results from a survey undertaken by this working group as an inventory of common practices in pilot and ATC selection.

The focus of this report is on pilot and air traffic controller selection. The working group initially planned to include other aviation professions, such as cabin crew and mechanics, in the report. However, the lack of literature on selection of these other professions led to a decision to focus on pilot and ATCO selection. Much of the information contained in this document pertains to selection in general and may be applied to other professions, but many of the topics addressed are too broad to be covered exhaustively in this report. Extensive reference lists have been provided for each chapter for those readers seeking a more in-depth understanding of the topic.

The order of the chapters in this report reflects the topics that may be considered in developing a selection system. Thus, a reader could read the entire report to understand more about how a selection system is developed and the issues to consider. Those readers who are interested only in a specific topic may elect to read only certain chapters. The chapters are independent of each other, although information is cross-referenced between the chapters.

The authors of this report would like to emphasize the fact they and EAAP do not endorse any specific commercially available selection product. The results of the EAAP Selection Survey presented in Appendix B include names of test providers because respondents were asked to indicate which tests they use. This question was included in the survey because EAAP often is approached by members interested in products used by other members.

2. Why is Selection Important in Aviation?

Jóhann Wium

When we select for a role, we are making a choice - accepting a portion of those who applied while rejecting others. While the rest of this paper will discuss the intricacies on how this choice is made, this chapter will focus on the reasons why we select at all.

Because we can make this choice without putting any thought into it, a better way of framing our question might rather be „Why should we spend time, resources and effort on selection?". While a choice may be unavoidable, why must it be complicated? Sometimes aviation organizations simply have difficulty filling roles and because of staffing shortages need employees to start immediately. Why not simply choose someone by convenience, choose someone at random, or train everyone who applies? Why is it important that we spend our limited time and resources determining who to choose?

2.1. Reasons for selection

There are a number of reasons why organizations and companies in aviation should – and in some cases must – spend time and resources when selecting personnel.

2.1.1. Competence

The simplest, most basic reason is that when you are choosing someone for a specific job it is necessary to find someone capable of performing said job. This capability is not automatically guaranteed. We spend time and resources on selection trying to find these competent and capable employees.

How difficult it is to find these individuals varies greatly. Competence can come in many forms and can present differently between jobs so there is no single method that always works in selection. Some jobs have historically been difficult to select for, with a high percentage of those selected being unable to fulfil the requirements of the role. Other jobs may have more flexible requirements, allowing more people to adapt to the role and perform it successfully. Whatever the case may be, inadvertently choosing someone who is incapable of performing the job will lead to negative consequences for both the organization and the individual selected.

The important question to ask is then "Does selection lead to choosing more capable employees?". What evidence exists that spending time, resources and effort in selection leads to choosing meaningfully "better" employees? How do we know that those we did not select wouldn't have performed just as well than those we selected?

The answer to that is simple. We know because it has been tried. In 1943 the US Army Air Force decided to see what the pass rate of pilot cadets would be if test score requirements were waived. Applicants were still tested, but instead of being rejected because of low scores, 1,311 cadets were chosen on first-come-first-served basis. Care was taken not to reveal that these cadets were in any way different from those selected with traditional methods. At the end of their training only 23% had successfully graduated from advanced flight training as opposed to 63.4% who graduated the following years after traditional selection. Of the 150 cadets that were admitted with the lowest test scores not a single one managed to graduate (Dubois, 1947).

While this is perhaps a unique example in aviation comparing selection and non-selection, there are numerous examples of selection improving graduation rates, for pilots as well as other roles. One example is that following a major re-design of selection for air traffic controllers in the early 80s, the Federal Aviation Administration increased the success rate for student air traffic controllers from 43% to 71% (Sells et al., 1984). Better selection of student air traffic controller has also shown to mean higher scores in training, higher supervisor ratings of performance, better employee retention and fewer disciplinary actions (Trites, 1961). Similarly, for pilots better selection has shown to reduce flying hours required in training. A study by Duke and Ree (1996) showed, for example, that no extra flying hours were needed for those scoring in the highest 40 percentiles of the initial selection procedure.

While today's research might focus on the small differences between different types of selection, there is more than sufficient evidence in the literature to support that selection in general can be used to choose significantly better and more capable aviation personnel.

2.1.2. Cost-effectiveness

Employees can affect the financial health of their employer in numerous ways, being capable of costing or saving them money through their actions. Some jobs even require expensive training where success is not guaranteed (e.g. pilots or air traffic controllers), with failure leading to the individual being removed from service, which costs the employer directly. Organizations might also try to select individuals with other positive characteristics (e.g. high motivation, interest, good job-fit, etc.), looking beyond basic competence. This is done in the belief that these characteristics benefit the company indirectly, eventually leading to better outcomes for the organization.

Therefore, spending money on appropriate selection for employees can result in saving the employer money. Those selected are more likely to pass training, are likely to stay longer, perform better and so on.

One simple way to conceptualize the potential savings is to look at the cost of training a single individual to completion. If the selection saves the organization from hiring one individual who fails out at the end of training this cost would represent the direct savings accomplished. This cost of training has been estimated for a number of roles. According to Mattock et al. (2019) the training cost of US Air Force fighter pilots ranges from $5,6 to $10,9 million dollars (depending on aircraft type). For air traffic controllers the savings are not quite as extreme but are still estimated at around $500,000 dollars each (Buck & Pierce, 2018). Finally, according to Goeters and Maschke (2004) the equivalent savings for selecting a successful airline pilot from licensed applicants could be around $150,000 (€125,000, adjusted for inflation).

Goeters and Maschke (2004) also include data on the cost of hiring and training failures and we can use that to calculate a simplified return on investment. For the selection system they describe for ab initio pilots, every Euro spent on selection returns more than €450 in savings to the company (giving it an ROI of 45,170%). For already licensed pilots the savings are less but still quite significant, or about €28 in savings per Euro spent on selection (giving an ROI of 2,871%).

While the above numbers show possible direct costs connected to selection there are other factors that can affect the cost/savings made by a good selection system with several different methodologies available to calculate this utility or cost-benefit analysis. For possible methods of estimating the utility of employee selection see for example Cascio, Boudreau and Fink (2019), Heintz (2004), or Zeidner and Johnson (1991).

2.1.3. Legal Requirements

Selection can be subject to domestic and even international regulations. Different countries have different laws and it is crucial that those responsible for selection are aware of the legal requirements that apply to them. Whatever the requirements may be, the selection process must fulfill these legal requirements for continued operation.

More on the legal requirements for selection can be found in chapters 5, 9 and 10 of this paper.

2.1.4. Safety

The importance of safety is undeniable in aviation and can affect how employees are selected. An aviation organization could be held responsible for the unsafe conduct of their employees and therefore spending time and effort selecting individuals who are less likely to exhibit such unsafe conduct is sensible. For many there is also a moral reason to pursue safety, as being safe is not just beneficial for the organization, but for society as a whole.

When the accident rate from the 1943 group of "non-selected" pilot cadets was analyzed it became evident that there was a marked difference between accident rates that could be predicted by the score on the traditional selection measures. Those who had scored in the lower 33% of test scores were three times more likely to have accidents compared to those who scored in the highest 33%. There were also 4 fatalities in the training of these 1,311 cadets and all four had scored in the lowest third (Flanagan, 1948).

The US military attempted to identify what made a pilot accident-prone for the next 60 years. Many of the early efforts examined the relation between training performance and operational accidents (Webb, 1955) but found no relation. Later studies attempted to isolate an "accident-prone" personality (for a review see Rodgers and Blanchard, 1993) but showed mixed results (see also Sanders, Hofmann, Hunt & Snow (1974) and its follow-up study by Sanders, Hofmann and Neese (1975), and King, Retzlaff and Orme (2001). Outside of aviation, meta-analysis has indicated that personality can predict unsafe behaviors (Beus, Dhanani and McCord, 2014) but there is no clear evidence to suggest an "accident-prone" personality exists for pilots or other aviation specific personnel.

It is possible that this is because aviation is a safety-focused industry with numerous safeguards against accidents already in place, and it will therefore be difficult to show the direct effects of specific employee attributes on accident rates. One method researchers have adopted is to look for the more common incident rates, that should go together with accident rates, and compare that to psychological measures. In such an attempt, Hunter and Stewart (2011) discovered that a previous history of hazardous event involvement correlates significantly with accident rate for pilots. With that data a clearer link has already been established between accidents and psychological attributes such as risk perception (Hunter, 2002) or locus of control (Hunter, 2006).

While the specific relationship between selection and safety remains elusive, it is a good practice to consider possible safety implications when looking at how employees are selected.

In practice, however, the lines between these specific reasons for selection are easily blurred. We select in order to choose competent, capable employees. Because they are competent, they are more likely to pass training, saving the company money compared to someone who was less competent. The legal requirements are there to ensure that those selected are less likely to cause accidents, thereby improving safety through better performance and reducing the likelihood of a costly incident. In the cost sensitive, regulation heavy and safety critical industry of aviation, the reasons why we select are highly interconnected and rarely are we selecting just for one of these reasons. It is more likely that all of them will apply.

2.2. Summary

On the whole, selection is important because it matters who we choose for a role. Although a single individual is less likely to have a dramatic effect on a large organization than a small one, it is always possible. A single employee can end up having significant and long-lasting effects on the organization he works for – whether they be positive or negative. Despite equipment and assets, organizations are ultimately composed of their employees, and their actions make the difference between stellar successes and dismal failures. This is why selection remains important, as choosing the right people can make all this difference.

2.3. References

Alkov, R. (1972). Are there accident-prone pilots? Paper presented at the Annual Meeting of the Human Factors Society. pp. 456-458.

Alkov, R. A., Borowsky, M. S., & Gaynor, J., A. (1983). Pilot error as a symptom of inadequate stress coping. Second International Symposium on Aviation Psychology. 401-405): The Ohio State University.

Beus, J. M., Dhanani, L. Y., & McCord, M. A. (2015). A meta-analysis of personality and workplace safety: addressing unanswered questions. The Journal of Applied Psychology, 100(2), 481–498. DOI: 10.1037/a0037916

Buck, J.L., and Pierce, L.G. (2018). Understanding the Air Traffic Control Field Training Process from the Perspective of the Developmental Controller. Report DOT/FAA/AM-18/13. Washington, DC: Federal Aviation Administration.

Cascio, W.F., Boudreau, J.W., and Fink, A.A. (2019). Investing in People: Financial Impacts of Human Resources Initiatives (3rd Edition). Alexandria, VA: Society for Human Resource Management.

Dubois, P.H. (Ed.) (1947). The Classification Program. Report No. 2. Army Air Forces Aviation Psychology Program Research Reports. Washington, DC: Army Air Forces.

Duke, A.P., & Ree, M.J. (1996). Better Candidates Fly Fewer Training Hours. Another Time Testing Pays Off. International Journal of Selection and Assessment, 4(3), 115-121.

Flanagan, J. C. (1948). The Aviation Psychology Program in the Army Air Forces Report No. 1. Washington, D.C.: U.S. Government Printing Office.

Goeters, K.M, and Maschke, P. (2004). Cost-Benefit Analysis of Pilot Selection: The Economic Value of Psychological Testing. In K.M. Goeters (Ed.) Aviation Psychology: Practice and Research (p. 203-208). Aldershot: Ashgate

Heintz, A. (2004). Cost-benefit Analysis in the Selection of Air Traffic Controllers. Presented at the 26th Conference of the European Association of Aviation Psychology (EAAP), Sesimbra, Portugal.

Hunter, D. R. (2002). Development of an aviation safety locus of control scale. Aviation, Space, and Environmental Medicine, 73, 1184–1188.

Hunter, D. R. (2006). Risk perception among general aviation pilots. International Journal of Aviation Psychology, 16, 135–144.

Hunter, D.R., and Stewart, J.E. (2011). Hazardous Events and Accident Involvement by Military and Civilian Pilots. The International Journal of Aviation Psychology, 21:2, 123-134. DOI: 10.1080/10508414.2011.556451

King, R. E., Retzlaff, P. D., & Orme, D. R. (2001). A comparison of US Air Force Pilot Psychological Baseline Information to safety outcomes (AFSC-TR-2001-0001). Retrieved from Kirtland Air Force Base, NM:

Mattock, M.G., Asch, B.J., Hosek, J., and Boito, M. (2019). The Relative Cost-Effectiveness of Retaining versus Accessing Air Force Pilots. Santa Monica, CA: RAND Corporation.

Rodgers, M.A., and Blanchard, R.E. (1993). Accident Proneness: A Research Review. Report DOT/FAA/AM-93/9. Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Sanders, M. G., Hofmann, M. A., Hunt, P. D., & Snow, A. C. (1974). Personality aspects of pilot-error accident involvement (75-1). Retrieved from Fort Rucker, AL

Sanders, M. G., Hofmann, M. A., & Neese, T. A. (1975). A cross validation study of the personality aspects of involvement in pilot-error accidents (75-15). Retrieved from Fort Rucker, AL

Sells, S.B., Daley, J.T., and Pickrel, E.W. (1984). Selection of Air Traffic Controllers. Report FAA-AM-84-2. Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Trites, D.K. (1961). Problems in Air Traffic Management: I. Longitudinal Prediction of Effectiveness of Air Traffic Controllers. Report FAA-AM-61-1. Oklahoma, OK: Federal Aviation Agency, Civil Aero-medical Research Institute, Research Branch.

Webb, W. B. (1955). Grades in training and subsequent accidents (55-4). Retrieved from Naval Air Station Pensacola, FL:

Zeidner, J., & Johnson, C. (1991). Classification Efficiency and Systems Design. Journal of the Washington Academy of Sciences, 81(2), 110-128. Retrieved January 26, 2021, from http://www.jstor.org/stable/24531452

3. Historical Overview of Pilot Selection

Diane Damos & Hans-Jürgen Hörmann

Pilot selection has a long history that began early in the 20th Century. In this chapter, we provide a broad overview of this effort. Over the last 100 years, thousands of studies on pilot selection have been conducted. It is impossible, therefore, to present an exhaustive review of each topic. Instead, we have attempted to include references that are representative of research conducted during a specific period on a specific topic.

Organizing the research chronologically makes it easier for the reader to see when specific research topics were introduced and how they developed over time. The careful reader will notice that some topics have been investigated continuously since the first decades of the 20th century, but few studies were conducted in any given decade. Others, in contrast, were extensively studied for a relatively short period, and then research on the topic was essentially stopped because the problem was assumed to have been solved or was no longer considered important.

Readers will note quickly that most of the references cited in this chapter pertain to military pilot selection. This is because militaries adopted aviation quickly for use in combat and needed trained pilots, but few pilots were available in the second decade of the 20th Century. Consequently, the military began ab initio (candidates with no flight experience) pilot selection. Pilot candidates needed careful selection to reduce training failures, which resulted in research to identify and assess relevant attributes. Additionally, militaries had and continue to have the financial resources and the candidate pool to conduct research on pilot selection. Civilian agencies, such as the Federal Aviation Administration in the United States (US) and the European Union Aviation Safety Agency (EASA) in Europe, have no mandate to be involved in pilot selection. The predominance of military selection articles is further exacerbated by the fact that civilian test developers, flying schools, and airlines frequently do not want their selection instruments and processes made public because of proprietary concerns.

Much research is published first as a conference proceeding, then as a report, and eventually as a journal article or book section. When two or three sources were available for a given study, the journal article or the book section is cited because they are the most easily accessible. Similarly, conference proceedings were preferred over technical reports. No selection instruments are discussed without published predictive validity data. The authors want to point out we encountered some of the same problems writing this chapter that investigators conducting meta-analysis encounter: Much research is not obtainable. Some of this “grey literature” are internal reports that were never meant to be published. Others were early government reports that no longer exist in the archives. This chapter, therefore, is not exhaustive.

3.1. 1910 to mid-1930s

Today, it is difficult to imagine developing a selection system for a position with no job analysis, few incumbents, and not even a job description. Because of the novelty of the pilot's tasks, that is exactly the challenge that psychologists faced in the second decade of the Twentieth Century. During this period, the need to select pilots was stimulated both by the high failure rate in training and the high cost of training. Accidents were relatively common (see Lamont & Sweetser, 1920, p. 27 for information on Canadian training accidents), resulting in a significant number of injuries, fatalities, and damage or loss of training aircraft, all of which drove up training costs. Pilot selection systems were developed in part to reduce accidents and training costs (see Chapter 2.1.2).

The decade from 1910 to 1920 saw the development of pilot selection systems by Britain, Canada, France, Italy, Germany, and the US. One of the most comprehensive sources of early pilot selection literature is a bibliography developed by Hoff and Fulton (1942), which includes entries for articles published in Russian, Polish, German, Chinese, Italian, Spanish, Portuguese, and Japanese. Dockeray and Isaacs (1921) give an overview of the development of the Italian, French, British, and American selection systems. Brief descriptions of the British system also are given by Anderson (1919) and Bailey (1999). Descriptions of the US selection system may be found in Damos (2007) and Stratton, McComas, Coover, and Bagby (1920). During this early period, the type of selection instruments employed varied widely from country to country. The British relied completely on interviews. The US exclusively used paper-and-pencil tests and inventories in its pilot selection battery although apparatus tests were under development. German investigators, in contrast, developed and employed apparatus tests in pilot selection; Stern (1919) and Moede (1930) used driving aptitude tests including an apparatus test that measured motor response patterns in reaction to audio-visual stimuli.

By 1920, most of the major attributes that today are considered important for a successful flying career had been identified. These include intelligence, perceptual speed, reaction time, motor coordination, personal history (biodata), and personality. From approximately 1920 to the late 1930s, interest in pilot selection declined in some countries, including the US and Britain (Bailey, 1999). Consequently, few new selection instruments were developed in either of these countries. One notable exception was the complex coordinator, a test of eye-hand-foot coordination

(Mashburn, 1934), which was developed in the US in the late 1920's. Other countries continued research on pilot selection, with two (Germany and Japan) adopting the complex coordinator quickly (Hopkins, 1944).

Although scheduled air carrier service began with British Airways in 1919, little information on civilian pilot selection is available until the late 1930s. In 1939, the US Civil Aeronautics Administration (CAA) began an ambitious program to interest young people in careers as civil aviation pilots (Viteles, 1945). Shortly after the effort began, the program received funding to develop selection instruments for civilian ab initio pilots and identify methods to improve civilian ab initio pilot training. Investigators concerned with selection developed a biodata form and tests of general intelligence, aviation knowledge, and mechanical comprehension. All these tests were predictive. Investigators studying training quickly identified the lack of standardized civilian testing and evaluation procedures as major problems. Consequently, the CAA developed a set of maneuvers to be used in pilot evaluation as well as a standardized rating form. The form was used by flight instructors to rate the overall competency of a student. This form proved to be a major advancement in pilot evaluation and was adopted quickly (with modifications) by a US airline and by the US Navy.

Despite the amount of research devoted to ab initio selection and training, very little research and development was performed during this period in the US on the selection of experienced airline pilots. One exception to this was Snow (1926) who claimed to have constructed a battery for airline pilots. He provides no information on how he identified the required attributes although he did state that the same types of abilities and personality traits were needed for anyone involved in transporting goods and people. All of the validation was conducted on taxicab drivers and streetcar operators. No subsequent reports were found indicating that Snow's battery was ever validated for pilots or used for pilot selection. No other attempts to develop selection instruments for experienced pilots were located for the US or any other country.

3.2. Mid 1930s through 1959

The period from the mid 1930s to the early-1940s saw a large increase in research on pilot selection. Hopkins (1944) provides brief descriptions of the Japanese and German pilot selection systems in 1935 and 1936. Bailey (1999) gives a succinct description of the development of the Royal Air Force (RAF) pilot selection systems. In Germany, Gestalt psychology had a major influence on pilot selection (see Kreipe, 1950, for a summary). Gestalt psychologists advocated selecting applicants based on their entire character and not just on specific skills. As one of the German pioneers in aviation psychology, Gerathewohl (1944, 1954) provided a detailed analysis of the psychological requirements of flying in instrument meteorological conditions. His work on the effects of acceleration forces on spatial orientation and motor coordination became seminal, especially in military pilot selection and training in Germany and in the US.

The US World War II effort is documented in a 19-volume series on aircrew selection and training, the Army Air Forces Aviation Psychology Program Research Reports (see Flanagan, 1948, for an overview). The research described in these volumes is unique in its sample sizes, which are often over 1000, and the number and types of selection instruments that were studied. A job analysis was conducted early in the development process to identify the attributes needed to pass all stages of flight training. This analysis identified a wide variety of cognitive, psychomotor, and personality attributes, as well as background information, that could potentially be used as predictors. Information on the development and validation of selection instruments assessing these attributes is found predominately in two volumes, Apparatus Tests (Melton, 1947) and Printed Classification Tests (Guilford & Lacey, 1947). The Army Air Forces research effort also included an unrestricted sample of 1152 men that was followed through all stages of flight training to determine the success rate of student pilots who have undergone only medical selection (DuBois, 1947, p. 187) (see Chapters 2.1.1 and 8.1.1, for more details). The success rate was 23%.

In the late 1940s, both the British RAF and the US Air Force decided to reduce the size and complexity of their pilot selection batteries. As noted by Bailey (1999), the RAF reduced the total number of tests in the battery but tested the same aptitudes. The US Air Force, in contrast, decided in 1953 to eliminate all psychomotor tests. This decision occurred because the testing apparatus was difficult to maintain, forcing the US Air Force to restrict the number of testing locations to those that could maintain the equipment adequately. This restriction resulted in bottlenecks (see Passey & McLaurin, 1966, for a discussion of testing issues) that disrupted the flow of candidates.

The development of selection batteries for experienced airline pilots lagged that of military pilots by many years. Perhaps the first systematic attempt to identify important attributes for airline pilot selection was performed by Gordon (1949). He began by surveying five major US airlines to determine how they selected their pilots. Gordon then constructed two matched groups of pilots. Pilots in the first group had had a successful career, whereas pilots in the second group had been dismissed from the airline before upgrading to captain. He then compared the groups' test scores on three selection instruments used by their respective airlines. Neither the intelligence test, the mechanical aptitude test, nor the Minnesota Multiphasic

Personality Inventory (MMPI) distinguished the two groups. Similarly, number of flight hours and the amount of prior training did not differentiate the two groups. Gordan then used the critical incidents technique to identify 21 tasks as critical for flight safety. He then recommended the development of selection instruments to assess the attributes required by the 21 tasks. The extent to which his findings and suggestions were incorporated in subsequent selection batteries is unknown.

The first selection system constructed for and used to select experienced airline pilots was developed by SAS (Scandinavian Airlines) in 1951 (Trankell, 1959). This system was based on a job analysis of SAS pilots, which identified 14 attributes that were important for success as an SAS pilot. These attributes were assessed using either standardized tests or direct observation by psychologists. The SAS selection system was so successful that only few changes in the selection process were made between 1951 and 1991 (Gordon, 1991).

3.3. 1960 through 1979

During this period, research focused on four areas: motor performance, multiple-task performance (timesharing), the use of simulators for selection, and personality. All four topics had been examined before and used in pilot selection systems but, for motor and multiple-task performance, the increasing capability and decreasing cost of dedicated computers opened new venues for task construction and performance measurement. The decreased cost of computers had a similar effect on the feasibility of using simulators for pilot selection. Personality, in contrast, benefited from advances in theory.

Although tests of motor performance had been dropped from the US Air Force's pilot selection battery in 1953, subsequent research demonstrated that paper-and-pencil tests could not assess the same skills (See Benel, 1976; Griffin & Koonce, 1996, for extensive reviews of psychomotor testing for military pilot selection). Consequently, the US Air Force re-examined two tests used in the Army Air Forces aircrew selection battery: the two-hand coordination test and the complex coordination test (McGrevy & Valentine, 1974). The two tests were updated to take advantage of modern electronics. Both were found to make significant contributions to the predictive validity of the battery and were re-instituted in the selection battery in 1993 (Carretta & Ree, 1993a; Carretta & Ree, 1993b; Griffin & Koonce, 1996). Much of the other research on motor performance during this period was concerned with manual tracking[1]. More specifically, new performance measures

[1] The apparatus tests developed as part of the Aviation Psychology Research Program included several tests that today are classified as tracking tasks. Such tasks have had predictive success from the early 1940s to the present

(Jex, McDonnell, & Phatak, 1966; Savage, Williges, & Williges, 1978; Wickens & Gopher, 1977) and new ways of adapting the difficulty of the tracking task were developed (Gopher, Williges, Williges, & Damos, 1975; Williges & Williges, 1978).

Interest in multiple-task performance was first evident in the 1971 study of dichotic listening by Gopher and Kahneman (1971). This was followed by a several studies examining the predictive validity of measures of multiple-task performance for both civilian and military pilots (Damos, 1978; Damos & Lintern, 1981; North & Gopher, 1976). These studies demonstrated promising results. However, multiple-task paradigms presented a host of practical problems involving performance measurement, practice, and the use of feedback. Much research during this period was concerned with these methodological issues (Fichtbauer, 1975; Griffin et al., 1979; North, Harris, & Owens, 1978) but the identified solutions were sometimes impractical. Additionally, the computer power available during this period was insufficient for operational selection batteries to include multiple-task tests.

The use of simulators as selection instruments was also explored during this time. Performance measurement was, and continues to be, a significant issue; simulators can produce vast amounts of data. Investigators during this period were faced with several challenges. The first was to identify which of the hundreds of parameters a simulator can generate should be used to assess specific aspects of performance. Two studies by Hill and Eddowes (1974) demonstrated this type of exploratory research. In the first study, they examined 326 parameters and over 2400 in the second study. Investigators also had to determine what type of measure—such as average error, standard deviation, root mean squared error, maximum deviation, etc.—best reflected performance on a given parameter (for discussions of this issue, see Fuller, Waag, & Martin, 1980; Hubbard, 1987; Hubbard, Rockway, & Waag, 1989). Additionally, for these types of measures, the investigators had to decide when to start or stop data collection for a given maneuver. For example, after a turn to straight and level flight, should data collection begin a few seconds into the straight and level segment or as soon as the wings are level?

The results of research on performance assessment were used to develop simulator-based selection tests. These tests typically consisted of a short syllabus of basic

---

and appear in a variety of forms in many pilot selection batteries. Consequently, they are worth describing briefly. At their most basic, tracking involves following something that varies (Kelley, 1968, p. 26). Tracking tasks are usually grouped into two categories: pursuit and compensatory. In a visual pursuit tracking task, the candidate manipulates a cursor with a control device to keep the cursor on a moving target. An example of this is fighter pilot trying to align a gunsight on another aircraft. Thus, in pursuit tracking the candidate sees the movement of the target independently of the movement of the cursor. In a compensatory tracking task, the target is fixed, and the cursor moves. The difference between the target and cursor is the error. In compensatory tracking, the candidate sees only the error and not the movement of the target or of the cursor independently. An example is the pilot attempting to hold a constant altitude of 5000 ft MSL Compensatory tracking tasks are usually considered to be more difficult than pursuit tracking tasks because the candidate does not know if the error results from a change in the target direction, speed, or acceleration.

aircraft maneuvers. The candidate's performance was measured on the maneuvers and, in some cases, the rate of improvement in aircraft control was also assessed. Long and Varney (1975) provide an early example of this approach. They developed a 5-hour syllabus of basic maneuvers and used an automated performance assessment system to determine how well a candidate learned basic aircraft control. Nine flight parameters were scored. Each parameter had a designated value ("fly a heading of 030") and a tolerance for each parameter was established. Absolute deviations from the designated value, the standard deviation, average time out of tolerance, frequency of out of tolerance excursions, and the mean amplitude of the out of tolerance excursions were analyzed. As early as 1960, the Swiss psychologist Franz Gubser published a description of the Instrument-Coordination Analyzer (ICA), a low-fidelity simulator of simple instrument-flight maneuvers (Gubser, 1960). The ICA combined objective measurement of the flight control task with behavioral observations of how the candidates coped with elevated stress levels. Modern versions of this test are still in use for ab initio pilot selection in Europe (e.g., Hörmann & Noser, 2018).

Personality research increased during this period, largely because of advances in theory and the development of new assessment instruments. The "Big Five" personality dimensions were discovered (Tupes & Christal, 1961) although their predictive validity for pilots was not yet evident. Kragh (1960) produced the first validity data for the Defense Mechanism Test (DMT), a projective personality test, for pilot selection. His early data showed low interrater reliability and moderate predictive validity to pass/fail from flight training but were promising enough to support continued research. The use of the Eysenck Personality Inventory for pilot selection began during this period (Jessup & Jessup, 1971). Because personality questionnaires for the general population had low predictive validities for ab-initio pilot training, Kirsch, Goeters, and Ewe (1975) developed the Temperament Structure Scales (TSS), a questionnaire reflecting some specific personality characteristics relevant for airline pilots[2].

3.4. 1980 through 1989

The 1980s were marked by four notable trends. First, the development of cognitive psychology led to the inclusion of information processing and cognitive tests in selection batteries to assess a broader range of attributes than in previous batteries.

[2] The TSS measures eleven personality characteristics, which had proven to be relevant in selecting pilot applicants. The eleven TSS-scales are Achievement Motivation, Rigidity, Mobility, Vitality, Extraversion, Dominance, Empathy, Aggressiveness, Instability, Spoiltness, and Openness (i.e., Social Desirability).

Second, the price of computers continued to decrease as the computing power and available storage increased. This allowed more widespread use of computerized tests. Additionally, authoring systems became available, which allowed the rapid development of tests by psychologists. Third, research on personality continued to evolve with significant resources devoted to identifying traits that were predictive of flight training performance. Fourth, women began to enter both military and civilian aviation in noticeable numbers.

As noted earlier, Gopher and Kahneman published a study in 1971 using dichotic listening, a cognitive test, for pilot selection but this article attracted little attention. The publication of the second article (Gopher, 1982) attracted a great deal of attention and resulted in a substantial research effort by the US Navy to validate dichotic listening for pilot selection (Griffin & Mosko, 1982). However, unlike the Israeli test reported by Gopher, the Navy's version included an updated version of the complex coordination test (Delaney, 1992; Griffin & McBride, 1986). This version of the test required the candidate to perform a one-dimensional tracking task using rudder pedals and a two-dimensional tracking task with a control stick. The dichotic listening test performed under multiple-task conditions was significantly correlated with measures of training performance.

Spatial ability tests were often included in pilot selection batteries in the 1980's. Perhaps the most common test of spatial processing was the rotated letters test. This test was based on work by Shepard and Metzler (1971). Most instantiations of this test for pilot selection purposes required the candidate to mentally rotate a letter or an abstract figure in two-dimensional space (Carretta, 1987; Damos & Gibb, 1986). Less frequently, the candidate had to rotate a three-dimensional figure in space (Gordon & Leighty, 1988; Roscoe & Corl, 1987). Another spatial test, the manikin test, also was used in several selection batteries (D'Arcy, 1987; Damos & Gibb, 1986).

The theory behind many information processing tests required accurately measured reaction time as a dependent measure. Including such tests in operational batteries became feasible during this period because of the increase in computer power with the decrease in cost. Cognitive tests frequently used in selection batteries included the Posner Letter Match Task (Carretta, 1987; Posner & Mitchell, 1967) and the Sternberg Item Recognition Test (Carretta, 1987; Sternberg, 1966). Multiple-task tests continued to be investigated and generally were found to increase the predictive validity of the selection system (Goeters, Hörmann, & Maschke, 1989; Griffin & McBride, 1986).

During the 1980s, computerized test batteries became operational. Some of these early batteries were developed for military pilot selection (Bartram, 1987; Carretta, 1987; Damos & Gibb, 1986) and others, for civilian pilot selection (Goeters et al., 1989; Roscoe & Corl, 1987). The WOMBAT (Roscoe & Corl, 1987) may have been the first PC-based pilot selection battery. Computerizing selection batteries resulted in more standardized test administration and objective test scoring. Scoring was often completed immediately, which allowed for greater throughput.

The changes in computer power and cost also resulted in improved simulator fidelity. As noted in the previous section, some organizations used simulators as testing devices for late-stage ab initio selection. This use of simulators continued throughout the 1980s. An example was the Canadian Air Force system, which consisted of five, one-hour lessons. Lessons were developed using a building-block approach, and candidates were scored automatically on their deviations from assigned parameters (James, 1985; Spinner, 1989). Other simulation-based selection systems administered the test after the candidates had some minimum amount of flight training (e.g., Fowler, 1981). Research also continued on performance scoring methods (Connelly & Shipley, 1982; Hubbard et al., 1989) during this period.

Advances in personality theory led to an increased interest in using personality measures to select pilots. Consequently, new types of personality tests were developed for pilot selection. For example, tests assessing risk taking (Carretta & Siem, 1988; Dale & Bartram, 1985; Dolgin, Shull, & Gibb, 1987; Shull & Dolgin, 1989) and self-confidence (Carretta & Siem, 1988) were developed and validated as pilot selection instruments. Research continued on other tests, such as the Eysenck Personality Inventory (Bartram & Dale, 1982). Aviation psychologists became aware of the five-factor model (McCrae & Costa, 1989) during this period although no "Big Five" pilot selection tests can be dated to this period. The DMT continued to be used during this decade although evidence of low predictive validity for Dutch and British military pilots began to accumulate (D'Arcy, 1987; Harsveld, 1991; Stoker, 1982). Additionally, research on Crew Resource Management (CRM) identified attitudes, communication skills, and other interpersonal behavior as critical for aviation safety. These findings stimulated research on specific personality attributes that are important for aviators (Gregorich, Helmreich, Wilhelm, & Chidester, 1989).

During this decade, women began to appear in appreciable numbers, both as military and as airline pilots. Because of the lack of prior research comparing the performance of the genders on selection instruments, investigators began examining male-female performance differences (e.g., Koonce & Berry, 1980). In most countries, the purpose of this research was to develop different tests norms for males and females. In the US, the Civil Rights Act of 1991 prohibits the use of different norms for selection tests. Consequently, the purpose of this research in the US was to identify selection instruments that needed to be revised to minimize male-female differences, or in extreme cases, discarded. Although investigation of male-female performance differences began in the 1980s, most of this research was conducted in the 1990s and will be discussed in the next section.

3.5. 1990 through 1999

The decade of the 1990s saw few new pilot selection instruments. Most innovations during this period concerned statistics and selection methodology, the application of existing pilot selection tests to new populations, and the use of assessment instruments from other domains. Perhaps the most important innovation during this decade was the acceptance and wide-spread use of meta-analysis. Meta-analysis allows psychologists to combine results from many studies to determine the relation between two variables. Meta-analysis was developed in the 1970s (Glass, 1976; Schmidt & Hunter, 1977) and became generally accepted in industrial/organizational psychology about a decade later. Aviation psychologists were able to use meta-analysis to accumulate the results from several studies to obtain, for example, a weighted average correlation (predictive validity) for a specific selection instrument (Campbell, Castaneda, & Pulos, 2010a; Damos, 1993; Lynch, 1991; Martinussen, 1996; Martinussen & Torjussen, 1998). They also used meta-analysis to identify moderator variables, such as aircraft type, the nationality of the testees or the time period when the data were collected (Hunter & Burke, 1994), that affected the predictive validity of the selection instrument under consideration. More details can be found in Chapter 8 on the validity of selection tests.

This decade also saw the development and use of assessment centers for pilot selection. Hörmann, Manzey, Maschke, and Pecena (1997) describe one example of this, which used the assessment center to evaluate both inter-personal skills and performance-related skills of ab initio and licensed pilots. With such behavior-based assessment techniques, candidates are rated by multiple observers on their behavioral preferences in work-related scenarios that reflect typical crew resource management tasks.

Another new area of research begun during this decade was concerned with selection of remote pilots for unmanned aerial systems (UAS or UAV). An example is Biggerstaff, Blower, Portman, and Chapman (1998), who report the development and initial validation of a selection battery for the UAV external pilot (the pilot responsible for takeoffs, landings and in-visual-range aircraft control).

Other areas of research, that began earlier, were continued. Personality inventories using the Big Five personality model were developed for pilot selection and became common during this time period (Anesgart & Callister, 1999; Callister, King, & Marsh, 1997; Shouksmith, 1993). Selecting based on personality traits to ensure good CRM behaviors also attracted interest during this period (Chidester, Helmreich, Gregorich, & Geis, 1991). Validation efforts continued on the 16 PF (Bartram, 1995) and the Eysenck Personality inventory (Bartram, 1995). The TSS were refined further and validated for direct-entry pilots (Hörmann & Maschke, 1996).

Research comparing male and female pilots continued. Again, in the US the purpose was to obtain norms (e.g., Carretta, 1997; King & Flynn, 1995) and to identify the selection instruments that showed the smallest difference between the sexes.

3.6. 2000 through 2009

This decade saw continued advances in quantitative methods pertaining to pilot selection. One method that received substantial attention was neural nets. As an alternative to classical clinical and statistical data aggregation in a selection context, artificial neural networks can be used to integrate available predictor information. Based on a preceding training phase, neural networks use classification algorithms to assign the candidates to categories (e.g., recommended or not recommended for training) by identifying typical performance patterns among the predictor test scores. Compared to linear regression methods, the neural networks approach automatically accounts for effects of nonlinearity, interactions, and suppression effects (Arendasy, Sommer, & Hergovich, 2007).

In a series of studies, Sommer and his colleagues compared the accuracy of prediction for neural nets versus common quantitative methods used in examining predictive validity. These included logistic regression (Sommer, Olbrich, & Arendasy, 2004) and discriminant analysis (Arendasy et al., 2007). In both studies, the neural nets were found to provide better predictive validity than the traditional method.

The most striking aspect of the research conducted during this decade was its emphasis on personality. The increased emphasis on personality may be attributed to two interrelated factors. The first was the development of new and different measures of personality that began in the previous decade. The second was the use of more comprehensive job analyses that included "soft" skills, like teamwork and communication. Research conducted during this period focused more on attributes that affect communication and teamwork in the cockpit than on specific personality traits that are related to objective performance measures.

Maschke and Goeters (2003) provide an example of this shift in emphasis. These authors had students in an airline flight school rate the importance of 76 different aptitudes with the Fleishman Job Analysis Survey (Fleishman, 1996; see also chapter 4.1). Their ratings were then compared to those of active airline pilots. Significant differences were found on 13 of the attributes with the airline pilots always rating the attribute more important than the students. The largest differences were found on the personality/social attributes. The airline pilot data were analyzed in more depth by Goeters, Maschke, and Eißfeldt (2004). They found that certain social skills, such as cooperation and communication, were rated as more relevant to an airline pilot's job than some cognitive abilities, such as selective attention and visualization, that are frequently used in selection batteries.

Measurement of social skills with self-report methods is problematic in a selection situation. Hörmann, Radke, and Hoeft (2007) administered two questionnaires assessing social skills and one personality inventory to ab initio candidates. These candidates also completed an assessment center evaluation. During the assessment center exercises, the candidates were scored on four dimensions of social competence. Two criterion measures were included in the study: successful completion of the selection process and a score estimating the candidates' likely success as an airline pilot. Only the assessment center scores and, to a lesser degree, the personality inventory predicted the candidate's likely success score. Self-reports of social skills provided no incremental validity for the criteria.

As noted by Maschke (2004), pilots with certain personality characteristics may be resistant to social skills training as taught in airlines' CRM courses. Adult personality is generally considered to be stable over time and resistant to change. Additionally, some investigators began questioning how well CRM could be trained. Consequently, Maschke advocates selecting candidates based on personality measures. That is, the candidate must not have personality traits that result in poor CRM behaviors and a resistance to CRM training. Nevertheless, as noted by Marschke, selecting based on personality measures has many issues including the lack of personality-related criteria, lower test reliability, possible non-linearity between the predictors and the criteria, and cultural differences that may affect test validity.

Few studies of cognitive abilities were conducted during this decade. One exception concerned dynamic spatial ability—the ability to project the movement of objects in two-dimensional space (D'Oliveira, 2003). This concept was not new; the US Army Air Forces had developed similar tests in their aviation psychology program late in 1943 using motion pictures to display movement (Gibson, 1947). The validation efforts were limited, and the predictive validities were moderate, at best. The use of dynamic spatial ability tests to predict pilot performance had been abandoned until D'Oliveira conducted her study. She displayed moving objects on computer screens. To increase her sample size, she combined student air traffic controllers with student pilots. The criterion was an overall measure of training performance scored on the same scale for each specialty. The participants also completed a battery of paper-and-pencil tests that included several tests of spatial ability. D'Oliveira does not show any results just for the pilots. Nevertheless, the dynamic spatial ability test did not correlate with the paper-and-pencil tests of spatial ability. The author found that dynamic spatial ability measures predicted training performance better than paper-and-pencil measures of spatial ability.

One other issue of major importance to aviation psychologists surfaced during this decade. Commercial sites offering practice on specific pilot selection tests became widespread. Such sites pose significant threats to test validity. Candidates using such sites can practice one or more selection tests as often as they want for as long as they want. The psychologist administering the selection system does not know which candidates practiced the tests, how much practice was received, and the practice schedule (massed versus distributed). Albers and Hoeft (2007) address this issue and recommend either developing practice-resistant tests or providing free, representative tests for practice to every applicant.

3.7. 2010 through 2019

Many studies published in this decade were concerned with the continued validation of existing pilot selection systems (e.g., Johnston and Catano (2013) for the Canadian Forces pilot selection system) or existing instruments (e.g. Caponeccia, Zheng, and Regan (2018) for the Wombat). Other studies were concerned either with exploring the predictive validity of new tests of attributes with demonstrated predictive validity or with using established tests on a new candidate population that had received limited examination in previous studies. Two of the attributes that received substantial attention during this decade were perceptual speed and spatial ability.

The predictive validity of a perceptual speed test for civilian student pilots was examined by Mekhail, Niemczyk, Ulrich, and Karp (2010). They found that the number of correct responses correlated significantly with time to solo, time to the private pilot certificate, and with the grade point average for 116 ab initio pilots.

Johnson, Barron, Carretta, and Rose (2017) performed a multi-study examination of the predictive validity of spatial ability and perceptual speed for US Air Force pilot candidates. In the first study, the authors obtained data from over 30,000 US Air Force officer candidates on five measures of spatial ability, two of perceptual speed, and five of quantitative and verbal ability. A principal components analysis using an orthogonal varimax rotation showed a 2-factor solution. One factor was defined by the quantitative and verbal ability tests, the other by the perceptual speed and spatial ability tests. The table of intercorrelations reveals a pattern similar to that of Lacey (1947), who found spatial and perceptual speed tests to be significantly correlated. The second study was a meta-analysis that compared traditional pilot selection instruments (measures of quantitative and verbal ability) to measures of perceptual speed and spatial ability. Criterion measures included pass/fail from flight training, flight grades, academic grades, and various performance composites. Perceptual speed had the highest mean weighted predictive validity across all criteria. In the third study, Johnson et al. examined the incremental predictive validity for spatial and perceptual speed tests to the US Air Force pilot selection battery. The criteria were both academic and flying grades. Although the spatial ability tests contributed little to the predictive validity of the existing battery to any criterion, the perceptual speed tests provided incremental validity to all criterion measures, especially to the measures of flying performance.

Some attributes, such as stress tolerance and motivation, have been difficult to measure. Traditionally, interviews and paper-and-pencil tests have been used to assess these attributes, but often these methods did not provide an accurate picture of the candidate. The German Armed Forces developed a series of academic and simulator lessons to assess such traits (Meierfrankenfeld, Gress, & Vorbach, 2015). Each simulator lesson had three parts: a demonstration by an instructor, an assisted attempt by the candidate, and a "solo" flight. During the solo flight, the candidate received no help from the instructor and was scored by the instructor and an aviation psychologist on attributes such as situational awareness, task saturation, concentration, and stress resistance. Objective performance measures were obtained from the simulator scoring system. To begin flight training, the candidate had to receive passing scores on both the academic and simulator lessons. The academic and simulator lessons demonstrated predictive validity and, after their adoption, failures in flight training because of flying deficiencies decreased from 10% to 5%.

One attribute that has received very little attention for pilot selection, is multi-cue probability learning. Matton, Raufaste, and Vautier (2013) administered a multi-cue probability learning test to 132 students who had been accepted for flight training at the French Air Transport Training School. The students were followed over 2.5 to 3 years of training. Students were grouped into two categories for analysis purposes: those who required no additional flight time to complete the training, and those who needed additional training or failed training and were eliminated from the program. Scores on the multi-cue probability task were grouped by quartile. The only significant difference in flight performance was between those students in the lowest quartile versus the other three quartiles.

Secondary school grades in technical subjects (science, physics, and mathematics) are sometimes used as cut scores for entrance into ab initio flight training. In other selection systems, school grades are included in the prediction equations with scores on selection instruments. Knowledge tests of these academic subjects may also be used as selection instruments. Zierke (2014) observed that the incremental validity provided by knowledge tests and secondary school grades is rarely examined. Consequently, he compared the predictive validity of knowledge tests and grades to that of cognitive tests from the DLR selection battery for ab initio selection at a major European airline. School grades and the knowledge tests had approximately the same predictive validity as the cognitive tests for pass/fail from flight training. The only exception was English: neither school grades for English classes nor the English test was predictive.

"Cognitive test" can refer to a wide range of selection instruments that test information processing, memory, intelligence, and executive functions. King et al. (2013) examined the predictive validity of three different cognitive batteries to training performance for US Air Force pilot candidates. The first test, the Multi-Attribute Battery (MAB), is a classic intelligence test battery. The MAB was administered to over 12,000 student pilots. In studies with sample sizes of this magnitude, many statistical tests are significant but not informative. That is, large sample sizes produce statistically significant correlations that are so low as to be practically unimportant. Therefore, the magnitude of the difference between the groups should be evaluated rather than just the significance of the statistical test. For all three batteries, the authors compared those candidates who passed the first phase of undergraduate flight training versus those who failed. Six of the ten subtests of the MAB showed significant between-group differences, with the group that passed scoring higher than those who did not. However, the differences were modest. Analyses conducted only on those who passed versus those who failed for flight deficiency reasons showed basically the same pattern: Candidates who passed scored higher on seven of the ten subtests than those who did not. Both groups had an average IQ of approximately 120. The second battery, the MicroCog, was developed mainly to assess clinical pathology. MicroCog has 18 subtests that are grouped into five domains, such as attention and memory. Each subtest produces multiple measures, and the domains produce summary measures. MicroCog was administered to over 5500 pilot candidates. The results were generally comparable to those from the MAB: Candidates who passed training scored higher than those who failed for any reason. The third test, CogScreen, assesses more basic abilities than the MAB and produces 65 scores. It was administered to more than 7000 pilot candidates. Again, those who passed flight training scored higher than those that did not. About half of the differences were statistically significant, but again, the differences were modest. The one exception to this was the attention-shifting test, where larger between-group differences were evident. King et al. concluded that none of the three batteries was more predictive than the current US Air Force pilot selection battery, and no subscale or cognitive function was more predictive than the others. The authors suggest that broad abilities, not specific abilities, predict pass/fail.

Personality continued to be a topic of interest during this decade. Campbell, Castaneda, and Pulos (2010b) conducted a meta-analysis on 26 studies that examined the relation between personality constructs and training outcome for British or American military pilot candidates. The studies used different instruments to assess different personality constructs. Consequently, only three constructs—neuroticism, extraversion, and anxiety—were assessed in enough studies to be included in the meta-analysis. Campbell et al. found that all three attributes were related to training outcome, although the effect sizes were small. Neuroticism was the best predictor of failure in flight training followed by anxiety, whereas extraversion was predictive of training success.

As noted earlier by Maschke (2004), personality assessments should be used to identify pilots who are resistant to the social skills training such as those taught in CRM courses. Hörmann and Goerke (2014) investigated the use of social competence questionnaires as a replacement for more time-consuming and expensive methods of assessing a pilot's social competency, such as interviews or assessment center exercises. The authors administered two tests of social competency as well as the TSS to 305 candidates for flight training at Lufthansa. The candidates completed the DLR selection battery. The study used three criteria. One criterion was accepted/rejected for flight training. The second was a score estimating how well the candidate would do in his/her career. This score ranged from 1 to 9 and was given only to those candidates who passed the selection process. The second criterion involved performance during flight training. Those students who completed the training with no irregularities were placed in one category, and those who failed or needed extra training time were placed in a second category. The TSS scores correlated with estimated career success and with performance in training. The social competency scores did not correlate with acceptance into flight school, estimated career success, or training performance. Nevertheless, one of the social competency questionnaires correlated with social behaviors observed during the assessment center exercises, and scores from the assessment center exercises correlated with training performance. Based on these findings, the authors suggest using a social skills questionnaire early in the selection process to eliminate weak candidates before administering expensive selection instruments, such as assessment center exercises.

A careful reading of this chapter reveals that no research has been presented for selecting civilian helicopter pilots. An exception to this is Dickens (2013), who constructed a selection system for North Sea helicopter pilots. The selection system consisted of an intelligence test, a Big Five personality inventory, a questionnaire on mental health, and an interview. The main finding of interest was that the candidates scores on the Big Five looked very similar to that of US Army helicopter pilots.

UAS pilot selection continued to be of interest during this decade. Cuevas, Kendrick, Zeigler, and Hamilton (2015) had 16 university students who were enrolled in a professional UAS curriculum fly a simulated surveillance mission that required two crewmembers. The crews' mission performance was evaluated on eight dimensions, including teamwork, spatial orientation, task management, and problem solving. Background information on the student's experience with manned and unmanned aircraft and his/her experience with team activities of all types was collected. The authors also obtained information on the student's experience playing four different categories of video games. Prior experience with manned aircraft was related to measures of teamwork, whereas the amount of time spent in team activities related to problem solving and task management. Only one of the four categories of video gaming experience correlated significantly with any measure of performance: time spent playing single-shooter games correlated with spatial orientation.

The increased need for UAV pilots during this period led investigators to explore expanding the use of selection batteries designed for fixed-wing pilots to UAV pilots. Carretta (2013) conducted such a study. US Air Force undergraduate UAV (URT) training is divided into three stages. Stage 1 is designed to provide a basic aviation skill set. It includes both academic and flight training. Stage 2 involves flight training in a simulator as well as academic training. Stage 3 provides academic training on tactics and theater operations. Carretta predicted pass/fail from URT. The same measures that predict success in fixed-wing training—performance on a battery of cognitive and psychomotor tests, score on the Air Force Officer Qualifying Test (AFOQT), and prior flight hours—predicted pass/fail from URT.

A similar study was conducted by Rose, Barron, Carretta, Arnold, and Howse (2014) on US Air Force UAV candidates undergoing the training described above. They used the same three measures as Carretta (2013) as predictors but included a personality inventory, the SDI+, that assesses the Big Five personality factors plus a sixth factor that assesses Machiavellianism. Predictive validity was determined for pass/fail from Stage1 and for academic grades, average check ride score, and average daily flying grade for Stage 1. Rose et al. found that the three measures used for US Air Force fixed-wing pilot selection—score on the AFOQT, score on the battery of cognitive and psychomotor tests, and prior flight hours—had good predictive validity to pass/fail from Stage 1 training and the academic average and average daily flying score for Stage 2. None of the personality measures predicted performance in Stage 1, and only Openness correlated significantly with any measure of performance in Stage 2.

Range restriction is a ubiquitous problem in pilot selection. Candidates who scored poorly on one or more of the selection instruments are not allowed to begin training. Those who begin training, therefore, are a restricted sample of the candidate cohort. Calculation of predictive validities on restricted samples typically underestimates the predictive validity in an unrestricted sample. The aviation psychologist, therefore, must correct for the restriction, using some statistical method. In some cases, the psychologist has insufficient data to meet all statistical assumptions and must approximate the required information

Only two studies have been conducted on an unrestricted population. The first was conducted by the US Army Air Forces Aviation Psychology Program during World War II and was discussed earlier (DuBois, 1947). In the more recent investigation (Hörmann, Noser, & Stelling, 2018), an unrestricted sample of 135 Swiss Air Force pilot candidates completed the DLR battery. The TSS was also administered. The candidates then completed a two-week flying training course consisting of ground school and flight training. The criterion was the final performance evaluation of the flight instructor. Investigators examined the predictive validity of the selection instruments in the unrestricted sample and in two subsamples that were created mimicking a selection system with strict criteria (38.5% accepted) and one with more lenient criteria (77.8% accepted). All raw correlations decreased from the unrestricted to the lenient sample. However, not all of them decreased from the unrestricted sample to the strict sample. Of the eleven significant correlations in the unrestricted sample, five were significant in the lenient sample. Five correlations were also significant in the strict sample, but they were not the same five as in the lenient sample. Although no correlation was significant in the strict sample that was not significant in the unrestricted group, some of the correlations were larger. These findings indicate that dealing with restricted samples may be more problematic than many investigators realize.

3.8. Summary

In this review, we attempted to present a broad overview of the important developments in pilot selection from its beginning in the second decade of the 20th century to the start of the 21st century. The review is structured by time period rather than by attribute, such as psychomotor performance or spatial ability. Our goal was to present studies that are representative of the research conducted during a specific period on a specific topic. We did not always achieve this goal: some research articles could not be retrieved., others were not available in English or German. This resulted in bias towards English-, and German-language publications that were produced in either Europe or North America.

By structuring the research by time period rather than attribute, we hoped to demonstrate that advances in pilot selection were a combination of the development of new theories, selection instruments, statistical methods, and technologies, such as the wide-spread availability of low-cost computers. This format should also allow the interested reader to trace the development of tests of specific attributes, such as psychomotor coordination, over time.

Even a cursory examination of this chapter shows that, over more than 100 years, psychologists have examined numerous attributes to identify those that can predict success in flight training. A variety of instruments have been developed for those attributes that are demonstrably predictive. Because of the wealth of data available, today an aviation psychologist has ample information to identify the best selection instruments for the selection system in question.

We do not suggest, however, no advances in pilot selection will be made in the future. New theories, new methods of assessment, and new statistical techniques will lead to new ways of thinking about some of the attributes currently used in selection batteries and will result in new, reliable, cost-effective selection instruments. Changes in the pilot's tasks may also require the assessment of new attributes and the development of their associated instruments. These developments will present new challenges and new opportunities to aviation psychologists.

3.9. References

Albers, F., & Hoeft, S. (2007). Practice effects on test-takers' Performance and Quality of Cognitive Ability Tests in Pilot Selection: A spatial ability test. In 2007 International Symposium on Aviation Psychology (pp. 1-6). Dayton, OH: Wright State University.

Anderson, H. G. (1919). The Selection of Candidates for Aviation. In The Medical and Surgical Aspects of Aviation (pp. 16-40). London: Henry Frowde Hodder & Stoughton.

Anesgart, M. N., & Callister, J. D. (1999). Predicting training success with the NEO: The use of logistic regression to determine the odds of completing a pilot's screening program. In Proceedings of the Tenth International Symposium on Aviation Psychology. Columbus, OH: The Ohio State University.

Arendasy, M., Sommer, M., & Hergovich, A. (2007). Statistical judgment formation in personnel selection, A study in military aviation psychology. Military Psychology, 19(2), 119-146.

Bailey, M. (1999). Evolution of aptitude testing in the RAF. Paper presented at the RTO HFM Workshop on Officer Selection, Monterey, CA (pp. 25-21; 25-27). http://ftp.rta.nato.int/public/PubFulltext/RTO/MP/RTO-MP-055/MP-055-25.pdf

Bartram, & Dale. (1982). Eysenck Personality Inventory as a selection test for military pilots. Journal of Occupational Psychology, 55, 287-296.

Bartram, D. (1987). The development of an automated testing system for pilot selection: The MICROPAT project. Applied Psychology: An International Review, 36, 279-298.

Bartram, D. (1995). Personality Factors in Pilot Selection: Validation of the Cathay Pacific Airways Selection Procedures. Eighth International Symposium on Aviation Psychology, The Ohio State University, pp. 1330-1335). The Ohio State University.

Benel, R. A. (1976). SAM Complex coordinator and multidimensional pursuit test: Annotated Bibliographies. Perceptual and Motor Skills, 42, 327-338.

Biggerstaff, S., Blower, D. J., Portman, C. A., & Chapman, A. D. (1998). The Development and Initial Validation of the Unmanned Aerial Vehicle (UAV)External Pilot Selection System (NAMRL-1398). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Callister, J. D., King, R. E., & Marsh, R. W. (1997). Using the NEO-PI-R to assess the personality of US Air Force pilots (AL/HR-TR-1997-0097). Brooks Air Force Base, TX: Armstrong Laboratory.

Campbell, J. S., Castaneda, M., & Pulos, S. (2010a). Meta-analysis of personality assessments as predictors of military aviation training success. International Journal of Aviation Psychology, 20(1), 92-109.

Campbell, J. S., Castaneda, M., & Pulos, S. (2010b). Meta-analysis of personality assessments as predictors of military aviation training success. International Journal of Applied Aviation Studies, 20(1), 92-109.

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

Carretta, T. R. (1987). Basic attributes (BAT) system: Development of an automated test battery for pilot selection: AFHRL Technical Report.

Carretta, T. R. (1997). Male-female performance on U.S. Air Force pilot selection tests. Aviation, Space, and Environmental Medicine, 68(9), 818-823.

Carretta, T. R. (2013). Predictive validity of pilot selection instruments for remotely piloted aircraft training outcomes. Aviation, Space and Environmental Medicine, 84(1), 47-53.

Carretta, T. R., & Ree, M. J. (1993a). Basic Attributes Test (BAT): Psychometric equating of a computer-based test. The International Journal of Aviation Psychology, 3, 189-201.

Carretta, T. R., & Ree, M. J. (1993b). Pilot Candidate Selection Method (PCSM): What makes it work? (AL-TP-1992-0063). Brooks AFB, TX: Armstrong Laboratory.

Carretta, T. R., & Siem, F. M. (1988). Personality, attitudes, and pilot training performance: Final analysis (AFHRL-TP-88-23). Brooks AFB, TX: Air Force Human Resources Laboratory.

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

Connelly, E. M., & Shipley, B. D. (1982). Development of sensitive performance measures for selection of crews for flight training. Paper presented at the Annual Meeting of the Human Factors Society. pp. 825-829.

Cuevas, H. M., Kendrick, K. M., Zeigler, Z. A., & Hamilton, D. J. (2015). Investigating UAS operator characteristics influencing mission success. Paper presented at the 18th International Symposium on Aviation Psychology, Dayton OH (pp. 117-122).

D'Arcy, S. H. (1987). Brief on aptitude testing in the RAF. RAF Biggin Hill: Royal Air Force.

D'Oliveira, T. C. (2003). Dynamic Spatial Ability & Ability to Coordinate Information: Predictive Contributions of Dynamic Spatial Ability and the Ability to Coordinate Information for Air Traffic Controller and Pilot Selection. International Journal of Applied Aviation Studies, 3(2), 227-241.

Dale, H. C. A., & Bartram, D. (1985). Micropat: A progress report on the validation of an automated test battery for pilot selection. Paper presented at the The 3rd Symposium on Aviation Psychology, Columbus, OH (pp. 533-540).

Damos, D. L. (1978). Residual attention as a predictor of pilot performance. Human Factors, 20, 435-440.

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

Damos, D. L. (2007). Foundations of military pilot selection systems: World War I (Technical Report No. 1210). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Damos, D. L., & Gibb, G. D. (1986). Development of a Computer-Based Naval Aviation Selection Test Battery (NAMRL 1319). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Damos, D. L., & Lintern, G. (1981). A comparison of single- and dual-tasks measures to predict simulator performance of beginning student pilots. Ergonomics, 24(9), 673-684.

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

Dickens, P. (2013). Looking for Chuck and PT: An evidence-based approach to assessing helicopter pilots. Paper presented at the 17th International Symposium on Aviation Psychology, Dayton OH (pp. 651-656).

Dockeray, F. C., & Isaacs, S. (1921). Psychological research in aviation in Italy, France, England, and the American Expeditionary Forces. Comparative Psychology, 1(2), 115-148.

Dolgin, D. L., Shull, R. N., & Gibb, G. D. (1987). Risk assessment and the prediction of student pilot performance. 4th International Symposium on Aviation Psychology, Columbus, OH (pp. 480-485). Columbus, OH: Department of Aviation, The Ohio State University.

DuBois, P. H. (Ed.) (1947). The classification program Report No. 2. Washington, D.C.: U.S. Government Printing Office.

Fichtbauer, S. (1975). Ein mobile Kleinsimulator fur Tracking- und Mehrfacharbeitssuche (A mobile mini simulator for tracking and multi-task work). Cologne, DL: German Aerospace Center.

Flanagan, J. C. (1948). The Aviation Psychology Program in the Army Air Forces Report No. 1. Washington, D.C.: U.S. Government Printing Office.

Fowler, B. (1981). The aircraft landing test: An information processing approach to pilot selection. Human Factors, 23(2), 129-137.

Fuller, J. H., Waag, W. L., & Martin, E. L. (1980). Advanced simulator for pilot training: Design for automated performance measurement system. (AFHRL-TR-79-57). Williams AFB, AZ 85224: Air Force Human Resources Laboratory.

Gerathewohl, S., J. (1944). Psychologische Untersuchungen zur Bindflugeignung. Zeitschrift für angewandte Psychologie, 66, 361.

Gerathewohl, S., J. (1954). Die Psychologie des Menschen im Flugzeug. München: J. A. Barth.

Gibson, J. J. (Ed.) (1947). Motion picture testing and research Report No. 7. Washington, DC: U.S. Government Printing Office.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Goeters, K.-M., Hörmann, H.-J., & Maschke, P. (1989). The DLR test system ab initio pilot selection. In Fifth International Symposium on Aviation Psychology (pp. 663-668). Columbus, OH: Department of Aviation, the Ohio State University.

Goeters, K.-M., Maschke, P., & Eisfeldt, H. (2004). Ability requirements in core aviation professions: Job analyses of airline pilots and air traffic controllers. In K.-M. Goeters (Ed.), Aviation Psychology: Practice and Research (pp. 99-119). Burlington, VT: Ashgate.

Gopher, D. (1982). A selective attention test as a predictor of success in flight training. Human Factors, 24, 173-184.

Gopher, D., & Kahneman, D. (1971). Individual differences in attention and the prediction of flight criteria. Perceptual and Motor Skills, 33, 1335-1342.

Gopher, D., Williges, R. C., Williges, B. H., & Damos, D. L. (1975). Varying the type and number of adaptive variables in continuous tracking. Journal of Motor Behavior, 7, 159-170.

Gordon, H. (1991). Selection of ab initio pilot candidates: The SAS system. In E. Farmer (Ed.), Human Resource Management in Aviation (Vol. 1, pp. 19-27). Aldershot, UK: Avebury Technical.

Gordon, H. W., & Leighty, R. (1988). Importance of specialized cognitive function in the selection of military pilots. Journal of Applied Psychology, 73(1), 38-45. Retrieved from 10.1037/0021-9010.73.1.38.

Gordon, T. (1949). The Airline Pilot's Job. Journal of Applied Psychology, 33(2), 122-131.

Gregorich, S., Helmreich, R. L., Wilhelm, J. A., & Chidester, T. (1989). Personality based clusters as predictors of aviator attitudes and performance. Paper presented at the Fifth International Symposium on Aviation Psychology, Columbus, OH.pp. 686-691.

Griffin, G. R., & Koonce, J. M. (1996). Review of Psychomotor Skills in Pilot Selection Research of the U. S. Military Services. International Journal of Aviation Psychology, 6(2), 125-147.

Griffin, G. R., & McBride, D. K. (1986). Multitask performance: Predicting success in naval aviation primary flight training (NAMRL-1316). Naval Air Station Pensacola, FL: Naval Aeromedical Research Laboratory.

Griffin, G. R., & Mosko, J. D. (1982). Preliminary Evaluation of Two Dichotic Listening Tasks as Predictors of Performance in Naval Aviation Undergraduate Pilot Training (NAMRL 1287). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Griffin, G. R., Mosko, J. D., Harris, S. D., Jones, T. N., North, R. A., & Owens, J. M. (1979). Psychometric Properties of Dichotic Listening and IMPACT Tests: Intercorrelations, Reliabilities, and Relationship to Naval Aviation Selection Tests. Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Gubser, F. (1960). Beiträge zur Psychologie des Instrumentenflugs (Contributions to the psychology of instrument flying). Winterthur, Switzerland P.G. Keller.

Guilford, J. P., & Lacey, J. I. (Eds.). (1947). Printed classification tests Report No. 5. Washington, D.C.: U.S. Government Printing Office.

Harsveld, M. (1991). The Defense Mechanism Test and success in flying training. In E. Farmer (Ed.), Human Resource Management in Aviation (Vol. 1, pp. 51-65). Aldershot, UK: Avebury Technical.

Hill, J. W., & Eddowes, E. E. (1974). Further development of automated GAT-1 performance measures. (AFHRL-TR-73-72). Williams AFB, AZ 85224: Flying Training Division, Air Force Human Resources Laboratory.

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

Hörmann, H.-J., Noser, P., & Stelling, D. (2018, September). A psychometrician's dream: Test validities in an unrestricted sample of future Swiss air force pilots. 33rd EAAP Conference, Dubrovnik, Croatia). Dubrovnik, Croatia.

Hörmann, H.-J., Radke, B., & Hoeft, S. (2007). Measurement of social competence in pilot selection. International Symposium on Aviation Psychology, Dayton, OH, pp. 263-267). Dayton, OH: Wright State University.

Hörmann, H.-J., Manzey, D., Maschke, P., & Pecena, Y. (1997). Behavior-oriented assessment of interpersonal skills in pilot selection concepts, methods, and empirical findings. Paper presented at the Ninth International Symposium on Aviation Psychology, The Ohio State University, Columbus, OH (pp. 1110-1115).

Hörmann, H.-J., & Maschke, P. (1996). On the relation between personality and job performance of airline pilots. International Journal of Aviation Psychology, 6(2), 171-178.

Hoff, E. C., & Fulton, J. F. (1942). A bibliography of Aviation Medicine. Springfield, IL: Charles C. Thomas.

Hopkins, P. (1944). Observations on army and air-force selection and classification procedures in Tokio, Budapest, and Berlin. Journal of Psychology: Interdisciplinary and Applied, 17, 31-37.

Hubbard, D. C. (1987). Inadequacy of root mean square error as a performance measure. Fourth International Symposium on Aviation Psychology (pp. 698-704): Department of Aviation, The Ohio State University.

Hubbard, D. C., Rockway, M. R., & Waag, W. L. (1989). Aircrew performance assessment. In R. S. Jensen (Ed.), Aviation Psychology (pp. 342-377). Brookfield: Gower.

Hunter, D. R., & Burke, E. F. (1994). Predicting Aircraft Pilot-Training Success: A Meta-Analysis of Published Research. International Journal of Aviation Psychology, 4(4), 297

James, J. A. (1985). Canadian automated pilot selection system. The 3rd Symposium on Aviation Psychology, Columbus, OH (pp. 513-519). Columbus, OH: Department of Aviation, The Ohio State University.

Jessup, G., & Jessup, H. (1971). Validity of the Eysenck Personality Inventory in Pilot Selection. Occupational Psychology, 45(2), 111-123.

Jex, H. R., McDonnell, J. D., & Phatak, A. V. (1966). A critical tracking task for manual control research. IEEE Transactions on Human Factors in Electronics, HFE-7(4), 138-145.

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. International Journal of Aerospace Psychology, 27(3-4), 109-120.

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

Kelley, C. R. (1968). Manual and automatic control New York: Wiley.

King, R., & Flynn, C. F. (1995). Defining and measuring the 'right stuff': Neuropsychiatrically enhanced flight screening (N-EFS). Aviation, Space, and Environmental Medicine, 66(10), 951-956.

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

Kirsch, H., Goeters, K.-M., & Ewe, R. (1975). Faktorenanalyse eines neuen mehrdimensionalen Personlichkeits-Fragebogens (Factor analysis of a new multidimesional personality questionnaire) (DFVLR-FB-75-20): Cologne: German Aerospace Center.

Koonce, J., & Berry. (1980). Comparison of males and females in the prediction of basic flight performance. Paper presented at the Psychology in the Department of Defense, pp. 612-615.

Kragh, U. (1960). The Defense Mechanism Test: A new method of diagnosis and personnel selection. Journal of Applied Psychology, 44(5), 303-309.

Kreipe, K. (1950). German aviation medicine World War II. Vol. 1 (Vol. 1). Washington, D.C.: Department of the Air Force.

Lacey, J. I. (1947). Perceptual tests. In J. P. G. a. J. I. Lacey (Ed.), Printed classification tests Report No. 5. Washington, D.C.: U.S. Government Printing Office.

Lamont, G., & Sweetser, A. (1920). Opportunities in Aviation. New York: Harper and Bothers.

Long, G. E., & Varney, N. C. (1975). Automated pilot aptitude measurement system (AFHRL-TR-75-58). Lackland AFB, TX: Air Force Human Resources Laboratory.

Lynch, W. E. (1991). A meta-analysis of pilot selection tests: Success and performance in pilot training. (AFIT/GSM/LSM/91S-44). Air Force Institute of Technology: Wright-Patterson AFB, OH

Martinussen, M. (1996). Psychological measures as predictors of pilot performance: A Meta-Analysis.

International Journal of Aviation Psychology, 6(1), 1-20.

Martinussen, M., & Torjussen, T. (1998). Pilot Selection in the Norwegian Air Force: A Validation and Meta-Analysis of the Test Battery. International Journal of Aviation Psychology, 8(1), 33-45.

Maschke, P. (2004). Personality evaluation of applicants in aviation. In K.-M. Goeters (Ed.), Aviation Psychology: Practice and Research (pp. 141-151). Burlington, VT: Ashgate.

Maschke, P., & Goeters, K.-M. (2003). Ab-initio flight training and airline pilot performance: Differences in job requirements. In G. Edkins & P. Phister (Eds.), Innovation and Consolidation in Aviation. Aldershot: Ashgate.

Mashburn, N. C. (1934). The complex coordinator as a performance test in the selection of military flying personnel. Journal of Aviation Medicine, 5, 145-154.

Matton, N., Raufaste, E., & Vautier, S. (2013). External validity of individual differences in multiple cue probability learning: The case of pilot training. Judgment and Decision Making, 8(5), 589-602.

McCrae, R. R., & Costa, P. T., Jr. (1989). More reasons to adopt the five-factor model. American Psychologist, 44(2), 451-452.

McGrevy, D., & Valentine, L., D., Jr. (1974). Validation of two aircrew psychomotor tests (AFHRL-TR-74-4). Brooks AFB: Personnel Research Division, Air Force Human Resources Laboratory.

Meierfrankenfeld, K., Gress, W., & Vorbach, t. (2015). Simulator-based assessment of flight-specific aptitudes in German Armed Forces aircrew selection Paper presented at the 18th International Symposium on Aviation Psychology, Dayton OH. pp. 620-623.

Mekhail, A., Niemczyk, M., Ulrich, J., & Karp, M. (2010). Using the Table Reading Test as an Indicator for Success in Pilot Training. Collegiate Aviation Review, 28(1), 101-114.

Melton, A. W. (1947). Apparatus tests Report No. 4. Washington, D.C.: U.S. Government Printing Office.

Moede, W. (1930). Lehrbuch der Psychotechnik (The textbook of psycho-technique). Berlin: springer.

North, R. A., & Gopher, D. (1976). Measures of Attention as Predictors of Flight Performance. Human Factors, 18, 1-14.

North, R. A., Harris, S. D., & Owens, J. M. (1978). Test-Retest Reliability of Individual Differences in Dual-Task Performance (NAMRL-1248). Naval Air Station, Pensacola, FL: Naval Aeromedical Research Laboratory.

Passey, G. E., & McLaurin, W. A. (1966). Perceptual-Psychomotor tests in aircrew selection: Historical review and advanced concepts. (PRL-TR-66-4). Lackland Air Force Base, TX: Personnel Research Laboratory.

Posner, & Mitchell. (1967). Chronmetric analysis of classification. Psychological Review, 74, 392-409.

Roscoe, S. N., & Corl, L. (1987). Wondrous original method for basic airmanship testing. Paper presented at the The Fourth International Symposium on Aviation Psychology, Columbus, OH. pp. 493-499.

Rose, M. R., Barron, L. G., Carretta, Arnold, & Howse. (2014). Early identification of unmanned aircraft pilots using measures of personality and aptitude. International Journal of Aviation Psychology, 24(1), 36-52.

Savage, R. E., Williges, R. C., & Williges, B. H. (1978). Cross-validation of regression equations to predict performance in a pursuit tracking task. Paper presented at the Proceedings of the Human Factors Society. pp. 369-372.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62(5), 529-540.

Shepard, R. N., & Metzler, j. (1971). Mental rotation of three dimensional objects. Science, 171, 701-703.

Shouksmith, G. (1993). Finding the right stuff: Personality studies of airline pilots. In R. S. Jensen &

D. Neumeister (Eds.), Proceedings of the Seventh International Symposium on Aviation Psychology (pp. 455-460). Columbus: Ohio State University.

Shull, R. N., & Dolgin, D. L. (1989). Personality and flight training performance. Paper presented at the Human Factors Society 33rd Annual Meeting. pp. 891-895.

Snow, A. J. (1926). Tests for transportation pilots. Journal of Applied Psychology, 10, 37-51.

Sommer, M., Olbrich, A., & Arendasy, M. (2004). Improvements in personnel selection with neural networks: A pilot study in the field of aviation psychology. International Journal of Aviation Psychology, 14(1), 103-115.

Spinner, B. (1989). Using the Canadian Automated Pilot Selection System to predict performance in primary flying training: Straight and level flight. Paper presented at the Fifth International Symposium on Aviation Psychology, Columbus, OH.

Stern, E. (1919). Ueber eine experimentell-psychologische Eignungsuntersuchung für Flugzeugführer. Schriften zur Psychologie der Berufseignung und des Wirtschaftslebens, 8, 78-95.

Sternberg, S. (1966). High speed scanning in human memory. Science 153, 652-654.

Stoker, P. (1982). An empirical investigation of the predictive validity of the defense mechanism test in the screening of fast-jet pilots for the Royal Air Force. Projective Psychology, 27(1), 7-12.

Stratton, G. M., McComas, H. C., Coover, J. E., & Bagby, E. (1920). Psychological tests for selecting aviators. Journal of Experimental Psychology, 3(6), 405-423.

Trankell, A. (1959). The psychologist as an instrument of prediction. Journal of Applied Psychology, 45(3), 170-175.

Tupes, E. C., & Christal, R. E. (1961). Recurrent Personality Factors Based on Trait Ratings. (ASD-TR-61-97). Lackland AFB, TX: Air Force Systems Command.

Viteles, M. S. (1945). The Aircraft Pilot: 5 Years of Research A Summary of Outcomes. Psychological Bulletin, 42(8), 489-526.

Wickens, C. D., & Gopher, D. (1977). Control theory measures of tracking as indices of attention allocation strategies. Human Factors, 19(4), 349-365.

Williges, R. C., & Williges, B. H. (1978). Critical variables in adaptive motor skills training. Human Factors, 20, 201-214.

Zierke, O. (2014). Predictive Validity of Knowledge Tests for Pilot Training Outcome. Aviation Psychology and Applied Human Factors 4(2), 98-105.

4. Historical Overview of Air Traffic Controller Selection

Jóhann Wium

When discussing the history of air traffic controller (ATCO) selection it is important to keep in mind that it does not follow a single, straight line. For approximately the first 70 years, pilot selection instruments and methods were developed predominantly by militaries while in contrast ATCO selection was in the hands of each country's air navigation service providers (ANSP's). Each country had its own selection process, and developments in one country did not necessarily influence others until much later. For the purposes of this chapter we will discuss the general evolution of the ATCO selection, research specific to ATCO selection and the occurrence of notable „first", i.e. publications that show the adoption of a particular approach that later became commonplace. When compared to pilot selection (see chapter 3), developers of ATCO selection systems were less likely to be early adopters of selection methods. ATCO selection followed pilot selection 5 – 10 years later.

In many ways the Federal Aviation Agency in the US led the development of ATCO selection from the beginning of published reports in the 60s. The FAA was in a unique position to do so because of the sheer number of air traffic controllers that they needed to select, the speed of ongoing technological implementation (Hilton and Sells, 1984) and the number of aviation psychology personnel at their disposal (Seifert, 1967). Prior to 1980, ATCO selection in Europe was largely done as part of pre-existing civil servant or military selection (Hatting, 1991), and it wasn't until the 90s that non-US sources were published in any number, many following EUROCONTROL's work to establish guidance material for the selection of ATCOs (EATCHIP, 1996).

However, no matter where the research took place, student attrition rates have been the focus all throughout ATCO selection history. This attrition, candidates not successfully completing ATCO training, cost the relevant agencies (ANSP's, militaries, etc.) time and money and made planning for staffing levels problematic (Broach, 2017). It is this focus on training success and the drive to reduce failure rates that seems to be the reason for continuous research and re-validation of selection methods. This echoes the development of pilot selection systems as a way to reduce training costs (Blower, 1997; Henmon, 1919; Shipley, 1984).

Unfortunately, a significant portion of ATCO selection research is lost to us today. Many studies were only published as internal reports while others were marked classified. Largely though this research was simply never published. It is therefore important to keep in mind that this historical review is pieced together from those articles that are available to us today and may not be a complete or exhaustive report of all significant events. Much less ATCO selection research has been published than pilot selection research, and oftentimes there is only one extant and available source. Fortunately, when research on ATCO selection was published it was often in the form of government reports, which were standardized and contained minutia that might not have passed peer review. Therefore, unlike the pilot selection chapter when there are multiple reports, technical reports are preferred over journal articles and conference proceedings, as they are the most accessible and extensive. Similar to the pilot selection chapter, journal articles are still given preference over conference proceedings as they often provide more standardized and detailed information.

A final note must be made regarding job analyses conducted on the role of ATCOs. A selection process should be based on a job analysis (particularly worker-based job analyses) and at times changes in selection methodology were made in conjunction with a job analysis. While the focus of this chapter is advances and developments in selection, for a historical overview of job analysis in air traffic control see Wium & Eaglestone (2022) and chapter 5 also covers a number of available and current job analyses for ATCOs.

4.1. 1922 through 1949

Air traffic control had its beginning with the formation of the International Commission on Air Navigation in 1922 by the League of Nations (Marien, 2019). As air traffic grew, so did the need for air traffic controllers but there are few historical references from these early years to describe how these individuals were chosen to take on this burgeoning role.

The very first individuals who controlled air traffic had not been selected for the role as such but were airport personnel who were in the right place at the right time. The title of world‘s „first air traffic controller“ has been given to either G.J.H. „Jimmy“ Jeffs, who worked at the first aerodrome tower at Croydon Airport (Historic Croydon Airport, 2021) in 1922, or Archie League who worked at the airfield in St. Louis (Federal Aviation Administration, 2021). These first air traffic controllers were airfield employees with some experience as pilots themselves, giving them the necessary experience to excel in this role.

4.2. 1950 through 1959

Civil aviation boomed in the in the late 1940's and 50's and those agencies responsible for air traffic control (e.g. the FAA in the US) made previous "ATC relevant" work experience a requirement for applicants. As such an applicant had to have some experience as a pilot, radio or radar operator (or even as an air traffic controller) to be eligible. Consequently, a large portion of those who were selected during those decades were from the military (Cobb, 1965). This approach of selecting individuals for their prerequisite knowledge and assumed basic abilities became the common method during this time and early on the number of vacancies was such that they could be filled solely using these applicants (Trites, 1961). This approach of looking at previous experience is still influential today as applicants with previous aviation experience are often thought to have some advantage over those who do not.

As time went by, however, the agencies could no longer fill their vacancies with qualified applicants and needed to select applicants with either less relevant aviation experience or with no previous aviation experience at all. In 1964 the FAA changed their entry requirements so that candidates without previous aviation experience became eligible for the position (Cobb and Nelson, 1974). These applicants became known as ab initio‘s (ab initio is Latin and means „from the start“), a term often used to refer to candidates who have no previous experience in the role they are being selected for.

4.3. 1960 through 1969

The very first tests intended to select air traffic controllers were developed in 1952 (see Brokaw, 1979). They were, however, never used as they were only intended to be used on individuals with previous ATC training. Following that project Brokaw suggested using psychological tests on all trainees and using their results to determine which tests predicted trainee performance. In 1959, a joint project by the FAA and United States Air Force showed that some of these psychological tests could adequately predict academic grades, instructor ratings and performance ratings of air traffic controller trainees (Brokaw, 1959). This prediction held even 5 years later in a follow-up study by Trites (Trites, 1961). This was followed by a number of FAA reports in the 1960s where the connection between specific tests, backgrounds, demographics and training success were measured and examined.

The "Problems in Air Traffic Management" series marked the first attempt to methodologically ascertain what factors were important for air traffic controller training success and performance on the job. The seven papers looked at: how well long-term prediction held from initial selection tests (instructor rating correlated r = .45 with supervisor rating 5 years later; tests correlated with same supervisor rating with r = .44; Trites, 1961); what tests best predicted performance in training (abstract, spatial, numerical, non-verbal analogies and job specific problems; Cobb, 1962); whether training-entry age mattered for success (Trites & Cobb, 1962) and how it interacted with ability and personality characteristics (there is an inverse relationship between training entry age and training success; Trites, 1964); whether previous job experience predicted training or job performance (experience is not an effective predictor alone for training success; Trites & Cobb, 1964b); and an analysis of different selection tests for terminal operations (added verbal abstract reasoning as a potential predictor; Cobb, 1965) and an analysis of available performance measures (instructor ratings were the most predictive of training success but a considerable halo factor was evident; Trites & Cobb, 1964c). The Problem series is the first large scale validation work done by the FAA and many of the topics covered there would be the subject of intense research in the following years.

Brokaw's original selection tests were a combination of civil service commission tests and commercially available off-the-shelf tests (Cobb, 1962) measuring a number of different attributes. While they did not follow a specific job analysis or model of desirable traits for ATCOs, Brokaw posited in a separate article that the ability to be an air traffic controller depended on a triad of factors: perceptual speed and accuracy, space relations, and reasoning and integration (Brokaw, 1979). In the years that followed the number and names of these factors changed but it has generally been accepted that cognitive abilities can predict performance (and training success) as an air traffic controller.

However, it was not only cognitive factors that were believed to be important to become an air traffic controller. Agencies also used personality measures to a differing extent when selecting air traffic controllers. In the US, however, ability tests and personality tests were used by different agencies as part of the selection of new student air traffic controllers.

The FAA had used personality tests in early research on ab initio selection (using both the California Psychological Inventory and psychologist ratings; Cobb, 1962) but discovered that adequate predictive modelling could be done without including any personality measures (Cobb, 1962; Cobb, 1965) leading to them being excluded from use in initial selection. Nonetheless, personality tests were mandated for ATCOs and had to be performed by aeromedical doctors as part of the medical examination to identify individuals unfit for duty (Cobb and Nelson, 1974). As part of this mandate in 1965, the 16PF was administered to over 10,000 ATCOs, identifying 1.2% of them for further examination. Following that examination, 31 ATCOs were relieved permanently or temporarily from duty (Convey, 1984). This marks the only occurrence where personality testing was used to disqualify job incumbents but personality testing continued to be performed on applicants since that date. For a summary of results from these early studies and comparison to the general population see Karson & O'Dell (1974a), Karson & O'Dell (1974b) and Convey (1984).

4.4. 1970 through 1979

One of the criticisms general ability tests faced for selecting ATCOs is that the tests used could be considerably different from the actual tasks the individuals are required to perform in the role (for a more detailed discussion on the comparison of ATC specific tests to general ability tests, see Eißfeldt, 2002). Therefore, in an effort to better select qualified candidates there was an interest in creating tests that mimicked certain aspects of a controller's job to see if the applicant had the requisite abilities to perform it. These are sometimes referred to as "ATC specific tests" and are intended to simulate specific tasks that are crucial to the role. The first such tests only captured a portion of the tasks or presented them in an obviously artificial form and can be considered to be a form of low-fidelity simulations.

The earliest attempts to capture this work-specific aptitude began in the 1950s with the development of the Air Traffic Problems tests, a pair of tests originally designed by the American Institute of Research and revised by the FAA (Trites & Cobb, 1964b). The tests required candidates to answer 30-60 simplified questions to determine if an aircraft is allowed to change altitude without violating a 5-minute time-separation rule (Cobb, 1962). However, as these were used during a period where applicants were required to have previous experience in aviation, they could also be considered early knowledge-based tests rather than specific work aptitude tests.

It wasn't until the 1970s that the FAA started designing work sample tests as we define them. Among them was the Motion Picture Test (Buckley and Beebe, 1972) or, as it became more commonly known, the Controller Decision Evaluation (CODE). This test used a motion picture projector to show simulated traffic and had the candidates predict conflicts as quickly and correctly as possible (Rock et al., 1981). The test had its limitations though, requiring considerable time, complicated scoring and equipment not generally available in Civil Service Commission testing sites (Pickrel & Dailey, 1979). As such the test was simplified and redone, containing fewer items and using slide projector, showing a new slide every 45 seconds instead of a motion picture and became the Multiplex Controller Aptitude Test (for details on the MCAT's development, see Dailey & Pickrel, 1977).

Another development was the Directional Heading Test (DHT). The DHT was a three-part test where applicants needed to quickly decipher directional information (cardinal directions and/or degrees on a compass) and determine headings and if a conflict would occur or not. The second part of the test required candidates to determine opposite headings and the third part required candidates to do so while under aural distraction (Cobb & Mathew, 1972). The test correlated .26 (.47 when corrected for restriction of range) with performance on the FAA's Screen (Schroeder, Dollar & Nye, 1990) suggesting potential incremental validity, but was difficult to administer and fell out of use soon after (Heil & Manning, 2001). At their core both the MCAT and the DHT were attempts to emulate a simple task done by air traffic controllers but turning it into an „applicant-friendly" exercise to be used in selection.

In addition to these low fidelity simulations, attempts were also made to estimate other cognitive abilities than classical verbal and numerical aptitude. Chiles and Smith developed a "code-lock" exercise (Chiles & Smith, 1971) intended to assess non-verbal problem-solving ability (the exercise correlated only .27 with other measures of non-verbal intelligence but highlighted several concerns when creating assessment exercises). Chiles, Jennings and West (1972 and 1974) adapted the Multiple Task Performance Battery from previous research on complex air crew performance, using it to assess time-sharing ability, but finding that while partly predictive the exercise did not provide enough incremental validity to be justified. While neither became standard practice in ATCO selection, both showed promise and were considered for further research.

During this period there is also notable interest in understanding the effects of demographic and biographical characteristics on training success. Four applicant characteristics were of special interest to researchers at the time: training-entry age, previous aviation experience, level of formal education and gender differences.

At the time, the discussion on training-entry age, and whether there should be an upper age limit for student ATCOs became highly controversial. In Trites and Cobb' (1964a) original study they state that older recruits, particularly those over 33 years old, are more likely to fail training or marginally pass. This was followed by a number of studies investigating the specific nuances, interrelationships and complexities of age effects. Age was found to have a negative effect on training performance, even when taking in ATCOs who had been trained elsewhere prior (with experienced candidates over 34 being three times more likely to fail than those who are younger and in an artificially "normalized" sample the correlation between age and training performance was r = -0.31; Cobb, Lay & Bourdet, 1971).

In a more detailed analysis done by VanDeventer, Taylor, Collins and Boone (1983) higher age was found to be detrimental to training success even within the

18-30 age range, and those results were replicated again in a study by Collins, Nye and Manning (with those 22 or younger having a 50% higher pass rate than those 28 or older; 1990). In addition, while previous ATC-relevant experience was beneficial it did not outweigh the age effects, although the authors point out that the difference is enough that it is possible to have different age limits for different applicant groups (with a 5-year difference between admission age for different groups found to be acceptable; VanDeventer & Baxter, 1984).

Previous aviation-relevant experience was also a topic of considerable research. Following Trites and Cobb' (1964b) results that not all types of aviation-relevant experience was beneficial for student ATCOs, more research followed to determine how much previous experience mattered. To complicate matters, positive effects of previous experience was confounded with negative effects of increased age (Cobb, Nelson & Mathews, 1974). A more thorough analysis of these effects was done by Cobb and Nelson (1974) where they determined that only previous direct ATC experience resulted in higher success rates while other previous aviation relevant experience did not, finding that 16% of candidates with ATC experience failed their academy training vs. 49% for pilots and 51% for candidates with no experience[3] and that candidates with previous pilot experience had the highest post-Academy attrition rates four years later (28-38% for pilot groups vs. 13-18% for non-pilots). In an attempt to better identify the applicability of previous experience the FAA developed the Occupational Knowledge Test (OKT), a test of ATC-relevant knowledge that applicants might possess (Lewis, 1978). According to a later analysis done on pass/fail rates, if the OKT had been used instead of stated prior experience for admission, pass rates could have been improved by up to 4% (Dailey & Pickrel, 1984).

Similarly, studies done on the effects of education found it to have no or negative effects (after correcting for both age and aviation experience) with regards to successfully completing training (Cobb, Young & Rizzuti, 1976). Gender was also found to have no effect on pass/fail rates or academy scores (Cobb, Mathews and Lay, 1972) and with little or no sex differences in personality measures (Karson & O'Dell, 1974).

4.5. 1980 through 1989

The 1980s saw the publication of two large works on ATCO selection and a considerable amount of validation research being done. While there were many contributing factors, one of the main factors was the dismissal of 11,400 American air

[3] Note that any pilot certificate was later found to be beneficial for training success in Pierce et al. (2013a).

traffic controllers following the strike of the Professional Air Traffic Controller Organization (PATCO), the then US union for air traffic controller in 1981. As this constituted close to 65% of FAA's operational workforce it became imperative to select quickly and in larger numbers than had been done before to ensure future staffing needs (Pickrel, 1984).

Both 'Selection of applicants for the air traffic control profession' (Rock, Dailey, Ozur, Boone and Pickrell, 1981) and 'Selection of air traffic controllers' (Sells, Dailey & Pickrell (eds.), 1984) provide exhaustive insight into FAA's selection procedures for air traffic controllers with the latter work being considerably more extensive and covering experience gained from selection after the PATCO strike. Both, however, present validation studies, information on test development, limitations on selection procedures, overview of ATCO job analyses, previous studies and research and represent the earliest cohesive works dedicated to ATCO selection.

Because of the unprecedented volume of selection in the US, this period of time is marked with refinement and research on those selection processes that were already in place. While the FAA made minor adjustments to their selection instruments there were fewer major changes during this period than in the years before with developments mostly focused on personality testing and work sample approaches.

While personality testing was still relegated to select-out in the US, identifying candidates who might not have the emotional tolerance to thrive in the role (with 48.9% high state-trait anxiety candidates passing Academy training vs. 61.7% low state-trait anxiety students; Collins, Schroeder and Nye, 1989), they did see an increased use outside of the US (see e.g. Helbing, 1964; Zeitsava & Togarev, 1985; Hilton and Sells, 1984; Convey, 1984) where they became a more integrated part of the selection process (see Hättig (1991) for a comparison of several countries).

As for work sample approaches, there was a marked increase in using computers to create low-fidelity simulations to help with the selection of air traffic controllers. Either these were early radar exercises that depended on time-sensitive decision making (e.g. for more on the development of the computerized Approach Control Task see Hättig, 1991 and Hättig, 1986) or a combined battery testing various, specific aspects related to air traffic control (e.g. for more on the Royal Air Force "computerization" of ATC test batteries see Hunter and Burke, 1986).

Alternatively, agencies began structuring their selection and training process in such a way as to consider the first part of their training as a practical demonstration of the capability of the candidate for the role of air traffic controller, often called a "screen". While this had been practiced as early as 1976 (Boone, 1984), FAA's "common screen" (Manning, Kegg and Collins, 1989) and "radar screen" (Boone, van Buskirk and Steen, 1980) were specifically intended as selection procedures rather than periods of training. Partly this was because the American legal system considered any part of a program that could exclude a person from further employment as selection rather than training (Della Rocco, Manning and Wing, 1990) but it was also a way to reduce costs, failing candidates after 9 weeks rather than considerably later, or up to 146 weeks (Manning, Kegg and Collins, 1988). The percentage of those who passed ATC training remained fairly constant after the introduction of the screen but with a higher portion failing earlier during the "screen". According to data from that time, an average of 38% failed during field training but after the screen's introduction, 30% failed in the "screen" with 8% failing during the following field training (Manning, Kegg and Collins, 1988), with similar numbers in the 80s after the workforce recovery following the PATCO strike (Della Rocco, Manning and Wing, 1990).

The screen had the added advantage of allowing researchers to correlate success in the "screen" with results from psychological tests (e.g. Schroeder, Dollar & Nye, 1990) instead of having to wait until the candidates finished their complete ATC

4.6. 1990 through 1999

In the 90s and 2000s, European air navigation service providers, who up until that point had been government agencies, were semi-privatized, often as limited companies or incorporated entities wholly owned by their respective government (International Civil Aviation Organisation, 2012; Arthur D. Little, 2006). With this, agencies that had previously selected air traffic controllers according to rules and regulations for government employee selection, were free to change their selection methodologies (for an overview of selection procedures for different European civil services, see European Union, 2020).

This coincided with the increasing ubiquity of personal computers and an increase in the use of low-fidelity simulations, either in the form of simulated radar exercises (e.g. Ackerman, 1992; Ackerman and Kanfer, 1993a; Ackerman and Kanfer, 1993b) or multi-part computerized batteries (e.g. Dover, Kaplan & Zami, 1993; Alava & Álvarez, 1999).

Such was the increased access to personal computers that it began to affect the results of ATCO selection procedures, with candidates with more computer game experience scoring higher on some tasks (with self-reported video game experience being significantly connected to better peformance on selection tests, accounting for between 1.3 - 9% of performane variance, depending on the specific task; Young, Broach & Farmer, 1997).

The decade also saw continuous innovation in selection tools. Included among those was a situational interview[4] using ATC-specific examples (Brehmer, 1995; Brehmer, 1998) that came out of the Swedish CAA's MRU project (Projekt Marknadsföring Rekrytering Urval och Utbildning) that was a concentrated effort to improve the selection and training of Swedish student ATCOs; the Dynamic Air Traffic Control (DAC) test, a low fidelity simulation exercise that also examined progressive learning potential (Eißfeldt, 1995); and the Computerized Assessment System for the Selection of ATCOs (CAS-ATCO) developed by Spain's ANSP (Alava & Álvarez, 1999). A considerable number of bespoke assessments were also made by various European agencies with a short description of a number of these available in EUROCONTROL's Information on Available and Emerging Selection Tests and Methods for ab initio Trainee Controllers (EATCHIP, 1997). Research on personality as a possible predictor continued and a number of studies examined it more closely. Schroeder et al (1993) found that the connection between scores on the NEO-PI and performance of the FAA's Screen was low but did increase the explained variance of their regression equations by 3%. Nye & Collins (1991) followed up on their previous research on state-trait anxiety and concluded that anxiety scores and MCAT performance together accounted for 29.5% of the variance for Academy final grades (see also Broach, Young and Schroeder, 1995).

It is during this time that cooperation between different countries on how best to select air traffic controllers began in earnest. As part of the European Air Traffic Harmonization and Integration Program attempts were made to determine "Best practices" in ATCO selection and create guidelines for ANSP's (EATCHIP, 1996). This was followed by a series of documents outlining specific aspects of the recruitment and selection of new ATCOs. While the original guidelines covered the entire selection process the reports that followed were more specific, covering such areas as marketing the role (EATMP, 2000; EATMP, 2003), recruitment, desirable candidate attributes (EATMP, 2001a), selection procedures (EATMP, 2001b) and recommendations for psychological tests, interviews and assessment centers (EATMP, 2002; EATCHIP, 1998). Appropriately for this increased cooperation, the first EUROCONTROL selection seminar was held in 1999 (EATMP, 1999), followed shortly by the FAA's international selection conference that same year. A further sign of this cooperative spirit was the publication of the first book devoted solely to the selection of air traffic controllers a few years afterwards, Staffing the ATM System (Eißfeldt, Heil, and Broach, 2002), written by authors collaborating from both Europe and the US.

[4] A situational interview is an interview where candidates are asked how they would respond in a hypothetical situation that could occur in the role. These answers are then compared to answers given by job incumbents and subject matter experts giving an indication of potential job performance (Brehmer, 1995).

4.7. 2000 through 2019

In the early 2000s, a number of European countries collaborated in the creation of FEAST, the First European Air Traffic Controller Selection Test, an integrated battery of multiple smaller tests designed to be used together to best select potential air traffic controllers (Rathje et al, 2003). Similarly, the FAA developed their own battery called the AT-SAT, following a similar approach of using multiple smaller tests (predominantly ability tests and low-fidelity simulations) as part of a single test battery (Ramos et al, 2001).

With the two selection batteries providing the first stage of applicant testing, this decade saw mostly the development and creation of assessments to use in other phases of the selection. These were both niche abilities and assessments that would take place in earlier or later phases in the selection process.

Examples of niche abilities are the development of a strip management test (Grasshoff, 2001), active listening tests (Bleckley & Wilkinson, 2015), focused attention (Fischer et al., 2016), eye movement conflict detection test (Gayraud et al., 2017), assessments of dynamic spatial ability (D'Oliviera, 2004), tests of field independence (Maliko-Abraham, 2001; Maliko-Abrahal, 2004), operational monitoring ability (Bruder et al., 2013), radar vectoring aptitude (Huston, Baldwin, Schulteis, 2017; Broach, 2019) or multi-tasking exercises (Lösch, Heintz and Kelava, 2012).

Studies on the effects of personality for ATCO student selection have had mixed results. Roe, Oprins and Geven (2012) performed a detailed analysis of the personality scores of over 200 student ATCOs admitted between the years 2003-2010 and found that different personality measures correlated with success at different stages of their training. Personality was found to add incremental validity to cognitive scores to predict progression in training (adding between 0.09 – 0.19 to validity measures). Luuk, Luuk and Aluoija (2009) found significant connection between personality and professional success for ATCOs but not the one they expected (they found that Extroversion had negative effects on becoming a successful ATCO but provided additional predictive power beyond cognitive measures).

Other studies focused not on ATCO student success but whether the personality profile of ATCO applicants or those who are selected to become students differed from the general population (e.g. Dollar & Schroeder (2004) for Myers-Briggs; Bleckley (2010 and 2013) for the Self-Description Inventory). Additionally, there was also research into non-cognitive attributes often connected to personality (such as emotional intelligence) and how they related to student ATCO success where researchers found that specific sections of an emotional intelligence test could be used to significantly increase prediction for graduation, in particular scores of stress tolerance, and reality testing (see Thompson, Chappelle & Goodman, 2014; Chappelle et al, 2015).

Researchers also began connecting personality assessments to other metrics than simply pass/fail. Oakes et al. (2001) looked at both cognitive and personality factors and their connection to skill acquisition and subsequent performance and King et al. (2008) used MMPI-2 scores to predict candidates seeking psychological help.

As researchers looked into other parts of the selection process, two phases in particular were of interest. Some studies looked at the final stages, looking critically at the assessment instruments of either assessment centers (e.g. Höft & Pecena (2004 and Pecena & Eschen-Léguedé (2006) studied the construct and criterion validity of assessment center and rater differences) or semi-structured interviews (e.g. Conzelmann & Keye's (2014) study on the incremental validity over cognitive ability tests). Others looked at what could be done to "pre-select", some even using assessments to immediately disqualify or reject candidates that were less likely to succeed training. Biodata was of renewed interest in 2010s and while such questionnaires had been in use earlier by some agencies (see e.g. Eißfeldt, 1999) the FAA's decision to implement it in 2013 lead to considerable scrutiny and a number of studies have been done since then to clarify their usage (see e.g. Pierce et al., 2013).

With the march of developing technology, new possibilities in assessment have also been studied and some have begun finding their way into ATCO selection. Research is being done on measuring a candidate's potential for stress resilience assessed through physiological response (Marko et al., 2019; Xidong et al., 2019) or through game-based assessments (Wium, 2021; Eaglestone, Arnold and Sligte, 2016; Boardman, 2017).

4.8. Summary

The intent of this chapter, similar to chapter 3 on pilot selection, was to chart the evolution and progression of selection methods for air traffic controllers. Presenting it in a linear fashion allows for a clearer picture of how changes and developments are connected and aids in a deeper understanding.

The history of air traffic controller selection allows the reader to see the increasing maturity and complexity of selection methods as well as research intended to either simplify the process or add incremental validity. Beginning with simpler knowledge and ability tests we see the gradual shift towards low-fidelity work samples and, as personal computers become more common, computerized test batteries. The desire to simplify the process can be seen in the development of both FAA's AT-SAT and EUROCONTROL's FEAST that allow for standardized testing packages and the increasing usage of biodata to screen candidates. Finally, the most recent steps are more focused on niche assessments of either specific abilities or using new technology – or both at the same time.

While it is difficult to say what future developments may influence how we select air traffic controllers it remains important to be aware of how we got to where we are before deciding where to go on to next. The inclusion of a history chapter such as this one will hopefully help practitioners when designing the next breakthrough in ATCO selection.

4.9. References

Ackerman, P.L. (1992). Predicting individual differences in complex skill acquisition: dynamics of ability determinants. Journal of Applied Psychology, 77(5), 598-614.

Ackerman, P.L., & Kanfer, R. (1993a). Development of an ability-based selection instrument for the MARC-Air traffic controller training program. Final Report. Minnesota Air Traffic Control Training Center.

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), 413-432.

Alava, J., & Álvarez, J. (1999). CAS – ATCOs, Computerised Assessment System for the Selection of ATCOs. In EATMP (1999), Proceedings of the First EUROCONTROL Selection Seminar – „Current and Required Future Selection Work and Methods in the ECAC Area". Report HUM.ET1.ST04.1000-REP-02. Brussels, Belgium: EUROCONTROL.

Arthur D. Little (2006). Understanding the Developments of Air Navigation Services in Europe. Industry Survey. Boston, US: Arthur D. Little.

Bleckley, M.K. (2010). Confirmatory Factor Analysis of the Self-Description Inventory (SDI+) in a Sample of United States of America (USA) Federal Aviation Administration Air Traffic Control Trainees. 29th Conference of the European Association of Aviation Psychology (EAAP), Budapest, Hungary.

Bleckley, M.K. (2013). The Self-Description Inventory+, Part 1: Factor structure and convergent validity analyses. Report No. DOT/FAA/AM-13/13. Washington, DC: FAA Office of Aerospace Medicine

Bleckley, M.K. and Wilkinson, C.L (2015). The Development of an Active Listening Test for Federal Aviation Administration Air Traffic Control Specialists. Proceedings 19th Triennial Congress of the IEA, Melbourne 9-14 August 2015

Boardman, E.C. (2017). Convergent and Predictive Validity of a Game-based Assessment in Air Traffic Control Recruitment. United Kingdoms: Arctic Shores.

Boone, J.O., van Buskirk, L., & Steen, J. (1980). The Federal Aviation Administration's Radar Training Facility and Employee Selection and Training. Report No. FAA-AM-80-15. Oklahoma, OK: Civil Aeromedical Institute

Boone, J.O. (1984). The FAA Air Traffic Controller Training Program with Emphasis on Student Assessment. In Sells, Dailey and Pickrel (Eds.) Selection of Air Traffic Controllers. Washington, DC: Federal Avaition Administration.

Brehmer, B. (1995). Psychometric analysis of a situational interview for the selection of air traffic controller candidates, Norrköping: Luftfartsverket, MRU delrapport No 18

Brehmer, B. (1998). The MRU situational interview Mark II. An analysis. Norrköping: Luftfartsverket, MRU delrapport No 29

Broach, D., Young, W. C., & Schroeder, D. J. (1995). Some personality and aptitude characteristics of air traffic control specialist trainees, Aviation, Space and Environmental Medicine 64(8), 711-6

Broach, D. (2017). Attrition in the U.S. Air Traffic Control Specialist (ATCS) Training: A Review of 50 Years of Data. 19th International Symposium of Aviation Psychology, p. 542-547.

Brokaw, L. D. (1959). School and job validation of selection measures for air traffic control training. Report W ADC-TN-59-39. Lackland Air Force Base, TX: Personnel Laboratory, United States Air Force.

Brokaw, L.D. (1979). Historical Overview of the Research and Development for Air Traffic Controller Selection. Paper presented at the 21st Annual Conference of the Military Testing Association. San Diego: Navy Personnel Research and Development Center.

Bruder, C., Grasshoff, D., Eißfeldt, H., & Gürlük, H. (2013). Aviator II – Simulator-based Research on Operational Monitoring and Decision Making for Human Operators in Future Aviation. Report DLR-FB-2013-04. Hamburg, Germany: Deutsches Zentrum für Luft- und Raumfahrt e. V. Institut für Luft- und Raumfahrtmedizin, Luft- und Raumfahrtpsychologie

Buckley, E. P. & Beebe, T. (1972). The Development of a Motion Picture Measurement Instrument for Aptitude for Air Traffic Control. Report No. FAA-RD-71-106. Washington, DC: FAA Systems Research and Development Services.

Chappelle, W., Thompson, W., Goodman, T., Bryan, C. & 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. 93-103. 10.1027/2192-0923/a000082.

Chiles, W.D., & Smith, R.C. (1971). A non-verbal technique for the assessment of general intellectual ability in selection of aviation personnel. (Report No. DOT/FAA/AM-71/28). Washington, DC: Federal Aviation Administration Office of Aviation Medicine

Chiles, W.D., Jennings, A.E., & West, G. (1972). Multiple Task Performance as a Predictor of the Potential of Air Traffic Controller Trainees. Report No. FAA-AM-72-5, Oklahoma, OK: FAA Civil Aeromedical Institute.

Chiles, W.D. & West, G. (1974). Multiple Task Performance as a Predictor of the Potential of Air Traffic Controller Trainees – A Follow-up Study. Report No. FAA-AM-74-10, Oklahoma, OK: FAA Civil Aeromedical Institute.

Cobb, B.B. (1965). Problems in Air Traffic Management V: Identification and Potential of Aptitude Test Measures for Selection of Tower Air Traffic Controller Trainees. Report No. FAA-AM-65-19. Oklahoma City, OK: Office of Aviation Medicine.

Cobb, B.B., & Matthews, J.J. (1972). A Proposed New Test for Aptitude Screening of Air Traffic Controller Applicants. Report No. FAA-AM-72-18. Washington, DC: FAA Office of Aviation Medicine.

Cobb, B.B. (1962). Problems in air traffic management: II. Prediction of success in air traffic controller school. Report No. DOT/FAA/AM-62-2. Oklahoma City, OK: FAA Civil Aeromedical Research Institute

Cobb, B.B. (1965). Problems in Air Traffic Management: V. Identification and Potential of Aptitude Test Measures for Selection of Tower Air Traffic Controller Trainees. Report No. FAA-AM-65-19. Oklahoma, OK: FAA Civil Aeromedical Institute

Cobb, B.B., & Mathew, J.J. (1972). A Proposed New Test for Aptitude Screening of Air Traffic Controller Applicants. Report FAA-AM-72-18. Oklahoma City, OK: FAA Civil Aeromedical Institute

Cobb, B.B., Carol, L., Young, B.S., & Rizzuti, B.L. (1976). Education as a factor in the selection of air traffic controller trainees. Report No. FAA-AM-76-6. Oklahoma, OK: FAA Civil Aeromedical Institute.

Cobb, B.B., Lay, C.D., & Bourdet, N. M. (1971). The Relationship Between Chronological Age and Aptitude Test Measures of Advanced-level Air Traffic Control Trainees. Report FAA-AM-71-36. Oklahoma, OK: FAA Civil Aeromedical Institute.

Cobb, B.B., Matthew, J.J., & Nelson, P.L. (1972). Attrition-Retention Rates of Air Traffic Control Trainees Recruited During 1960-1963 and 1968-1970. Report No. FAA-AM-72-33. Washington, DC: FAA Office of Aviation Medicine.

Cobb, B.B., Mathews, J.J., & Lay, C.D. (1972). A comparative study of female and male air traffic controller trainees. Report No. FAA-AM-72-22. Oklahoma, OK: FAA Civil Aeromedical Institute.

Cobbs, B.B., Nelson, P.L., & Mathews, J.J. (1973). Relationships between age, ATC experience, and job ratings of terminal area traffic controllers. Report No. FAA-AM-73-7. Oklahoma, OK: FAA Civil Aeromedical Institute.

Cobb, B.B., & Nelson, B.A. (1974). Aircraft-pilot and other pre-employment experience as factors in the selection of air traffic controller trainees. Report No. FAA-AM-74-8. Oklahoma, OK: FAA Civil Aeromedical Institute.

Convey, J. J. (1984). Personality Assessment of ATC Applicants. In Sells, Dailey and Pickrel (Eds.) Selection of Air Traffic Controllers. Washington, DC: Federal Aviation Administration.

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, 240–252.

Collins, W. C., Schroeder, D. J., & Nye, L. G. (1989). Relationship of anxiety scores to academy and field training performance of air traffic control specialists. Report No. DOT-FAA-AM-89-7. Washington, DC: FAA Office of Aviation Medicine.

Collins, W.E., Manning, C.A, & Nye, L.G. (1990). Studies of Poststrike Air Traffic Control Specialist Trainees: III, Changes in Demographic Characteristics of Academy Entrants and Biodemographic Predictors of Success in Air Traffic Controller Selection and Academy Screening. Report No. DOT/FAA/AM-90/4. Washington, DC: FAA Civil Aeromedical Institute

Dailey, J.T., & Pickrel, E.W. (1977). Development of New Selection Tests for Air Traffic Controllers. Report No. FAA-AM-77-25. Washington, DC: Federal Aviation Administration.

Dailey, J.T., & Pickrel, E.W. (1984). Development of the Air Traffic Controller Occupational Test. In Sells, Dailey and Pickrel (Eds.) Selection of Air Traffic Controllers. Washington, DC: Federal Aviation Administration.

Della Rocco, P.S., Manning, C.A., & Wing, H. (1990), Selection of Air Traffic Controllers for Automated Systems. Applications from Today's Research. In Wise, Hopkin and Smith (ed.), Automation and System Issues in Air Traffic Control. NATO ASI Series F, Volume 73.

D'Oliveira, T. (2004). Predictive Contributions of Dynamic Spatial Ability and the Ability to Coordinate Information for Air Traffic Controller and Pilot Selection. 25th Conference of the European Association of Aviation Psychology (EAAP), Zegrze, Polan.

Dover, S., Kaplan, E., & Zami, Y. (1993). Developing a selection system for Air Traffic Controllers on the low fidelity task-simulation-based approach to selection system development. Technical manual. Human Resources International Ltd.

Eaglestone, J., Arnold, A.W.J., & Sligte, I.G. (2016). The Use of Cognitive Games in ATC Selection. 32nd Conference of the European Association of Aviation Psychology (EAAP), Cascais, Portugal.

Eißfeldt, H. (1995). The Dynamic Air Traffic Control Test (DAC). In N. Johnston; R. Fuller & N. McDonald (Eds.), Aviation Psychology: Training and Selection. Proceedings of the 21st Conference of the European Association for Aviation Psychology Vol. 2 (pp. 130-136). Cambridge, United Kingdom: Avebury Aviation.

Eißfeldt, H. (1999). The use of biographical data in the pre-selection of ab inition ATC applicants. In Proceedings of the First Eurocontrol Selection Seminar. Report No. HUM.ET1.ST04.1000-REP-02. Brussels, Belgium: EUROCONTROL

Eißfeldt, H., Heil, M.C., & Broach, D. (2002). Staffing the ATM System: The Selection of Air Traffic Controllers. Aldershot, Ashgate.

EATCHIP (1996). Guidelines for Selection Procedures and Tests for Ab Initio Trainee Controllers. Report No. HUM.ET1.ST04.1000-GUI-01. Brussels, Belgium: EUROCONTROL

EATCHIP (1997). Information on Available and Emerging Selection Tests and Methods for ab initio Trainee Controllers. Report No. HUM.ET1.ST04.1000-REP-01. Brussels, Belgium: EUROCONTROL

EATCHIP (1998). Selection Tests, Interviews and Assessment Centres for Ab Initio Trainee Controllers: Technical Supplement. Report HUM.ST04,1000-GUI-03-02. Brussels: EUROCONTROL

EATMP (1999). Proceedings of the First EUROCONTROL Selection Seminar – "Current and Required Future Selection Work and Methods in the ECAC Area". Report No. HUM.ET1.ST04.1000-REP-02. Brussels, Belgium: EUROCONTROL

EATMP (2000). Marketing and Communication of the ATCO Job – Phase 1 – Study. Report HRS/MSP-002-REP-02. Brussels: EUROCONTROL

EATMP (2001a). Characteristics of Recruitment and Pre-selection of Ab Initio Trainee Controllers (Revised). Report HRS/MSP-002-GUI-02. Brussels: EUROCONTROL

EATMP (2001b). Guidelines for Selection Procedures and Tests for Ab Initio Trainee Controllers (Revised). Report HRS/MSP-002-GUI-01. Brussels: EUROCONTROL

EATMP (2002). Selection Tests, Interviews and Assessment Centres for Ab Initio Trainee Controllers: Guidelines for Implementation (Revised). Report HRS/MSP-002-GUI-03-01. Brussels: EUROCONTROL

EATMP (2003). Marketing and Communication of the ATCO Job - Phase 2 Study: The ATCO Job as Perceived by its Applicants. HRS/MSP-002-REP-03. Brussels: EUROCONTROL

Eißfeldt, H. (1999c). The use of biographical data in the pre-selection of ab initio ATC applicants. Proceedings of the First EUROCONTROL Selection Seminar - Current and Required Future Selection Work and Methods in the ECAC Area. Report No. HUM.ET1.ST04.1000-REP-02, pp. 23-34. Brussels, Belgium: EUROCONTROL.

Eißfeldt, H. (2002). ATC-Specific Tests. In Eißfeldt, Heil and Broach (Eds.) Staffing the ATM System: The Selection of Air Traffic Controllers. Pages 85-96.

European Union (2020). The Comparative Analysis of the Areas of the Civil Service Act in Selected European Countries. Prague, Czech Republic: Operational Programme Employment, European Social Fund.

Federal Aviation Administration (2021). First Air Traffic Controller Remembered. Retrieved February 22, 2021, from https://www.faa.gov/news/communications/controller_remembered/

Fischer, P., Abram, Y., Pillon, J., Choisay, F., Mourtay-Saussaye, M., Becker, C., & Duvillanrd-Monternier, S. (2016). Focused Attention for ATC applicants in the French Air Force. 32nd Conference of the European Association of Aviation Psychology (EAAP), Cascais, Portugal.

Gayrud, K., Eißfeldt, H., Hasse, C., Pannasch, S. (2017). On the Development of the Eye Movement Conflict Detection Test. 14th Conference on Psychological Assessment and Testing.

Geven, Es., Oprins, E., Veldhuijzen, E., & Roe, R.A. (2008). Using personality scales in ATC selection: Predictive validities for ATC task performance in Initial Training. 28th Conference of the European Association of Aviation Psychology (EAAP), Valencia, Spain.

Grasshoff, D. (2001). Der Strip Display Management Test (SDM). Internal report. Hamburg, Germany: DLR

Hättig, H.J. (1986). Simulation Based Testing: A New Approach in Selecting ATC Applicants. Proceedings of the 28th Annual International Military Testing Association, New London, Connecticut.

Hättig, H.J. (1991). Selection of Air Traffic Control Cadets. In R. Galand, and A.D. Mangelsdorff (Eds.), Handbook of Military Psychology (p. 115-129). New York, NY: John Wiley & Sons.

Heil, M.C., & Manning, C.A. (2001). Documentation of Validity for the AT-SAT Computerized Test Battery, Volume II. Report DOT/FAA/AM-01/6. Oklahoma City, OK: FAA Civil Aeromedical Institute

Helbing, J.C. (1964). Results of the selection of RNAF Air traffic and fighter control officers. In A. Cassie, S. D. Foldtema, & J. B. Parry (Eds.), Aviation Psychology. The Hague: Mouton.

Hilton, T.F., & Sells, S.B. (1984). Air Traffic Controller Selection in the United States and Other Countries: An International Overview. In S.B. Sells, J.T. Dailey and E.W. Pickrel (Eds.), Selection of Air Traffic Controllers. Washington, DC: FAA Office of Aviation Medicine

Historic Croydon Airport (2021). Genesis of Air Traffic Control. Retrieved February 22, 2021, from https://www.historiccroydonairport.org.uk/interesting-topics/air-traffic-control/

Höft, S. & Pecena, Y. (2004). Behaviour-oriented evaluation of aviation personnel: an assessment center approach. In Goeters, K.-M. (Ed.), Aviation Psychology: Practice and Research (pp. 153-170). Hampshire, England: Ashgate Publishing Ltd.

Hunter, D.R., & Burke, E.F. (1987). Computer-based Selection Testing in the Royal Air Force. Behavior Research Methods, Instruments & Computers, vol 19 (2), p. 243-245.

International Civil Aviation Organization (2012). Document 9980. Manual on Privatization in the Provision of Airports and Air Navigation Services (1st edition). Montreal, Canada: International Civil Aviation Organization.

Karson, S., & O'Dell (1974a). Personality makeup of the American air traffic controller. Aerospace Medicine, 45, 1001-1007

Karson, S., & O'Dell (1974a). Personality differences between male and female air traffic controller applicants. Aerospace Medicine, 45, 596-598

King, R.E., Schroeder, D.J., Manning, C.A., Retzlaff, P.D., & Williams, C.A. (2008). Screening air traffic control specialists for psychopathology using the Minnesota Multiphasic Personality Inventory-2. Report DOT/FAA/AM-08/13. Washington, DC: FAA Office of Aerospace Medicine.

Lösch, T., Heintz, A., & Kelava, A. (2012). Development of a Basic Multitasking Test for Operational ATM Staff Based on Multidimensional Item Response Theory. 30th Conference of the European Association of Aviation Psychology (EAAP), Villasimius, Sardinia.

Luuk, K., Luuk, A. & Aluoja, A. (2009). Predicting Professional Success of Air Traffic Control Personnel from Their Personality Profile at Admission to Ab Initio Training. International Journal of Aviation Psychology, 19. 235-251.

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, 523–548.

Maliko-Abraham, H.A. (2001). A Cross Sectional Study of Field Independence and United States Military Air Traffic Control Specialists. International Symposium on Aviation Psychology

Maliko-Abraham, H.A. (2004). A Study of Field Independence and United States Military Air Traffic Controllers. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting.

Manning, C.A., Kegg, P.S., & Collins, W.E. (1988). Studies of Poststrike Air Traffic Control Specialist Trainees: II. Selection and Screening Programs.

Manning, C.A., Kegg, P.S., & Collins, W.E. (1989). Selection and Screening Programs for Air Traffic Control. In Jensen, R.S. (ed.) Aviation Psychology. Gower Press, USA.

Marien, P. (2019). Counting Down To 100 Years of ATC. The Controller, vol 58 (2).

Nolan, M.S. (2009). Fundamentals of Air Traffic Control (5th edition).

Nye, L.G., & Collins, W.E. (1991). Some Personality Characteristics of Air Traffic Control Specialist Trainees: Interactions of Personality and Aptitude Test Scores with FAA Academy Success and Career Expectations. Report DOT/FAA/AM-91/8. Oklahoma City, OK: FAA Civil Aeromedical Institute

Pecena, Y. & Eschen-Léguedé, S. (2006). Assessment-Centre for Air Traffic Controller Selection: Assessor Training (ATR), Development and First Evaluation. 27th Conference of the European Association of Aviation Psychology (EAAP), Potsdam, Germany.

Pierce, L., Broach, D., Bleckley, M.K., & Byrne, C.L. (2013a). Using Biodata to Select Air Traffic Controllers. Report No. DOT/FAA/AM-14/8. Washington, DC: FAA Office of Aerospace Medicine

Pierce, L.G. Williams, C.A., Broach, D., & Bleckley, M.K. (2013b). Assessing prior experiences in the selection of air traffic control specialists. Report No. DOT/FAA/AM-13/7. Washington, DC: FAA Office of Aerospace Medicine.

Pickrel, E.W., & Dailey, J.T. (1979). Development of the Multiplex Controller Aptitude Test. Paper presented at the 21st Annual Conference of the Military Testing Association. San Diego: Navy Personnel Research and Development Center. p 60-67

Pickrel, E.W. (1984). Adjustments in the Air Traffic Service Following the PATCO Strike. In In S.B. Sells, J.T. Dailey and E.W. Pickrel (Eds.), Selection of Air Traffic Controllers. Washington, DC: FAA Office of Aviation Medicine

Ramos, R., Heil, M., & Manning, C. A. (Eds.) (2001a). Documentation of the Validity of the AT-SAT Computerized Test battery: Volume I. Report No. DOT/FAA/AM-01/5. Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Rathje, H., Golany, Z., & Eißfeldt, H. (2004). Pan-European Selection Test Battery for Air Traffic Control Applicants. In K. Goeters (Ed.), Aviation Psychology: Practice and Research (p. 171 – 201). Aldershot: Ashgate

Rock, D. B., Dailey, J. T., Ozur, H., Boone, J. O. & Pickrel, E. W. (1981). Selection of Applicants for the Air Traffic Controller Occupation. Report No. FAA/AM/82/11. Washington, DC: FAA Office of Aviation Medicine.

Roe, R.A., Oprins, E., & Geven, E. (2012). The Predictive Validity of Personality Tests in Air Traffic Controller Selection. 30th Conference of the European Association of Aviation Psychology (EAAP), Villasimius, Sardinia.

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. Report No. DOT/FAA/AM-90/8. Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Schroeder, D. J., Broach, D., & Young, W. C. (1993). Contribution of personality to the prediction of success in initial air traffic control specialist training. Report No. DOT/FAA/AM93/4. Washington, DC: FAA Office of Aviation Medicine

Seifert, R. (1967). Report on the 7th Conference for Aviation Psychology. The Controller, vol 6(4), p.16.

Sells, S.B., Dailey, J.T., & Pickrel, E.W. (Eds.) (1984). Selection of Air Traffic Controllers. Report No. FAA-AM-84-2. Washington, DC: Office of Aviation Medicine.

Thompson, W., Chappelle, W., & Goodman, T. (2014). Psychological Predictors of Air Traffic Control Training Outcomes. 31st Conference of the European Association of Aviation Psychology (EAAP), Valletta, Malta.

Trites, D.K. (1961). Problems in Air Traffic Management I. Longitudinal Prediction of Effectiveness of Air Traffic Controllers. Report No. DOT/FAA/AM-61-1. Oklahoma City, OK: FAA Civil Aeromedical Research Institute

Trites, D. K., & Cobbs, B. B. (1962). Problems in Air Traffic Management: III. Implications of. Training-entry Age for Training and Job Performance of Air Traffic Control Specialists. Report No. DOT/FAA/AM-62-3. Oklahoma City, OK: FAA Civil Aeromedical Research Institute

Trites, D.K., & Cobbs, B.B. (1964a). 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. Report No. DOT/FAA/AM-63-31. Oklahoma City, OK: FAA Civil Aeromedical Research Institute

Trites, D. K. (1964b). Problems in Air Traffic Management: VI. Interaction of Training Entry Age with Intellectual and Personality Characteristics of Air Traffic Control Specialists. Report No. DO/FAA/AM-65-21. Oklahoma City, OK: FAA Civil Aeromedical Research Institute

Trites, D. K., & Cobbs, B. B. (1964c). Problems in Air Traffic Management: VII. Job and Training Performance of Air Traffic Control Specialists - Measurements, Structure and Prediction. Report DOT/FAA/AM-65-22. . Oklahoma City, OK: FAA Civil Aeromedical Research Institute

VanDeventer, A.D., Taylor, D.K., Collins, W.E., & Boone, J.O. (1983). Three studies of biographical factors associated with success in air traffic control specialist screen/training at the FAA Academy. Report No. FAA-AM-83-6. Washington, DC: FAA Civil Aeromedical Institute.

VanDeventer, A.D., & Baxter, N.E. (1984). Age and Performance in Air Traffic Control Specialist Training. In VanDeventer, Collins, Manning, Taylor and Baxter (Eds.) Studies of Poststrike Air Traffic Control Specialist Trainees: I. Age, Biographic Factors, and Selection Test Performance Related to Academy Training Success. Report No. FAA-AM-84-6. Washington, DC: FAA Civil Aeromedical Institute.

Wium, J. (2021). Validating Stress Tolerance Exercise with Game-Based Assessment and Stress Checklists. 2021 International Symposium on Aviation Psychology. 13-18. https://corescholar.libraries.wright.edu/isap_2021/3

Wium, J., & Eaglestone, J. (2022). Job Analysis for Air Traffic Control: A Review of the Literature from 1949 to 2020. Aviation Psychology and Applied Human Factors, Volume 11 (1)

Young, W.C., Broach, D., & Farmer, W.L. (1997). The effects of video game experience on computer-based air traffic control specialist Air Traffic Scenario Test scores. Report DOT/FAA/AM-97/4. Washington, DC: Office of Aviation Medicine.

Zaitseva, E.N., & Tokarev, V.F. (1985). System Developed for Psychological Selection of AT Controllers. ICAO Bulletin, vol. 40 (10).

5. What Do We Want to Select & Assess?

Hans-Jürgen Hörmann

What we need to assess when selecting aviation personnel is determined by a job analysis and, in some cases, by additional regulatory requirements. Job analysis is not a single technique but rather a process of gathering and organising job-relevant information ranging from a review of critical incidents, identification of cognitive functions in job tasks, observation of work activities, to interviews and surveys with subject matter experts (SMEs).

This information helps to better understand how a job is done and which specific requirements should be considered for a sustainable match of human capabilities and the job tasks at hand. The goal of a job analysis is to determine a profile of personal attributes, which job incumbents should possess to accomplish their job tasks or roles successfully. These attributes are usually weighted by importance or criticality and can be visualized in form of a job requirements profile. Morgeson, Brannick and Levine (2020) provide a comprehensive overview of the diversity of job analysis methods. This chapter describes some commonly used approaches and findings in selection of aviation personnel.

5.1. Job Analysis Methods

The history of job analysis is described in a book chapter by Wilson (2014). According to Wilson, modern job analysis emerged in the early 20th century. The most significant contributions were published during the "golden era of job analysis" from 1941 to 1980. Examples are Flanagan's Critical Incident Technique (1948, 1954), McCormick's Position Analysis Questionnaire (1972, 1989), Christal developed the task inventory method of job analysis (1974), and Fleishman with his ability-oriented job-analysis and his ability taxonomy (1975, 1984). From a pragmatic point of view, Broach, Schroeder and Gildea (2019) summarized four important steps of a job analysis for personnel selection, which are in their opinion:

- review available information on the target job in order to derive distinct job tasks and/or KSAOs (KSAO stands for knowledge, skills, abilities, and other characteristics)
- build a survey using these identified job tasks and KSAOs and have them be assessed by SMEs (Subject Matter Experts), for example in terms of criticality and frequency for successful performance

- link each KSAO item to successful performance on the identified job tasks
- document the results to identify the most relevant attributes to be included in the selection program

Well-researched methods in occupational psychology to support this process are:

5.1.1. Critical Incident Technique (CIT)

According to Flanagan (1954) the CIT consists of collecting incidents of extreme behaviour, either outstanding or poor. The rationale is that extremes can be identified more easily than average everyday behaviour. Interviews, observations, and classification methods can be applied in this identification process. Based on the collected incidents critical performance requirements have to be defined in the sense of KSAOs. In this stage, job analysts have to be familiar with the specific occupation. Otherwise, intentions, principles, and standards of job activities cannot be interpreted appropriately. CIT has been successfully used for example to establish the critical requirements for US Air Force officers (Guilford and Lacey, 1947; Flanagan, 1948) and for ATCOs in Europe (Eißfeldt, 1988; Hansen, 1999; Reeb & Gabauer, 2016).

5.1.2. Cognitive Task Analysis (CTA)

Advanced automation levels in the workplace have led to increased cognitive demands for operators. As a powerful set of tools, CTA was developed especially for the system-design and training area in the 1980s and 1990s to delineate mental processes and skills needed to perform job tasks at an expert level (Seamster, Redding & Kaempf, 2017). Compared to traditional task analysis, CTA describes job tasks not on the behavioural level but in terms of cognitive structures and processes that underlie job expertise.

There are three primary facets of CTA: (1) knowledge elicitation, (2) data analysis, and (3) knowledge representation (Crandall, Klein and Hoffman, 2006). Since classical CTA can be very resource consuming, Applied Cognitive Tasks Analysis (ATCA) was developed as a more streamlined approach (Militello and Hutton, 1998; Seamster et al., 2017). Crandall et al. (2006) provide guidance through the complete set of tools. EUROCONTROL (1997) provided a model describing the main cognitive components and processes in air traffic control, which could serve to further refine cognitive task analysis methods for ATCOs.

5.1.3. Fleishman Job-Analysis Survey (F-JAS)

Based on a taxonomy of human performance, Edwin Fleishman developed a sophisticated survey, which can link job requirements primarily to specific human abilities for performing job tasks. In its original version, F-JAS contained fifty-two ability scales covering cognitive, sensory-perceptual, psychomotor, and physical domains of human performance (Fleishman and Quaintance, 1984). In a later stage, twenty-one interpersonal scales (e.g., dependability, social sensitivity) were added to the F-JAS-2 (Fleishman, 1996).

The F-JAS system provides clear definitions for each of the abilities and seven-point carefully anchored rating scales. On these rating scales, SMEs judge to what extent each ability is required for job (task) performance. Usually, job analysis data are collected and catalogued for the entire position, but it is also possible to focus on certain job tasks or on different levels of experience (e.g., novice vs expert). In The Handbook of Human Abilities (Fleishman and Reilly, 1992) the authors provide specifications for tests that measure each ability as well as examples of publicly available tests to measure these abilities.

5.1.4. Occupational Information Network (O*NET)

The research by Fleishman and his colleagues provided a foundation for the ability requirements subsequently included in O*NET, a database developed by the U.S. Department of Labor (Peterson et al, 1999). O*NET is not itself an analysis technique as the other described methods. However, it's database provides an easy-to-read wider range of relevant job information, which can be as useful as a single survey. O*NET scores cover cognitive, interpersonal, and physical skill requirements, as well as working conditions, and are mostly derived from survey data of large samples of job incumbents.

The information provided by O*NET (currently for over 1100 occupations) include a) the tasks and behaviours involved in work activities as well as b) the conditions under which work is performed, and c) the requirements imposed on job-holders in the form of experience (e.g., skills and expertise) and basic individual attributes (e.g., abilities, interests, and personality characteristics). This information is continuously updated for conditions in the U.S. In May 2021 the latest database version 25.3 had been released. International studies have shown that job information from O*NET is likely to transpose well across countries (Taylor, Li, Shi, and Borman, 2008). However, it seems still recommendable to adjust this information by experts for the local conditions in the respective organization.

This list of job analysis methods is certainly not exhaustive. Further classical methods are for example, the Repertory Grid Technique (e.g., Andersen, 1990; Hermans & Mulder, 1998) or the Position Analysis Questionnaire (McCormick et al., 1989). The book authored by Morgeson et al. (2020) could be consulted for more details.

5.2. Regulatory Requirements

Both ATCOs and pilots must fulfil several legal and/or quasi-legal requirements to get their medical exam class 1 or class 3 certificate issued or renewed (EASA, 2019; EUROCONTROL, 2006). These criteria should assure mental health by excluding problems with stress-coping, interpersonal behaviour as well as relationship issues and overt personality disorders. The criteria should be considered when executing a psychological evaluation.

More specific is regulation No 1042, which EASA issued in 2018. It requires aircraft operators to ensure that all flight crew members pass a psychological assessment before commencing line flying. This should prevent that psychological attributes or suitability of the flight crew in respect to the work environment could negatively interfere with the safe operation of the aircraft. According to the respective ANNEX III to this regulation a job analysis has to be conducted which details the complexity and the challenges of the operational environment that the flight crew is likely to be exposed to. The following assessment criteria should at least be included: (1) cognitive abilities, (2) personality traits, (3) operational and professional competencies, and (4) social competencies in accordance with crew resource management (CRM) principles.

5.3. Findings from Job Analyses for ATCOs

More than 60 studies to date have analysed different job positions in ATC. Broach (2002) described requirements for ATCOs in the US in the light of technological changes. He applied FAA's Job Analysis Information Base (JAIB, e.g., Knapp, Morath, Quartetti and Ramos, 1998), which links job tasks with KSAOs in a matrix format. Broach included ATCO tasks such as situation monitoring, resolving aircraft conflicts, managing air-traffic sequences, planning flight routes, assessing weather impact, managing sector resources, and responding to emergencies. These tasks were crosslinked to 15 KSAO categories and then evaluated in SME workshops. The top-five KSAOs were (1) conscientiousness, (2) memory, (3) attention, (4) heuristic reasoning, and (5) communication. The findings were updated in 2013 facing developments within the Next Generation Air Transportation System ("NextGen") (Broach, 2013). He found increasing importance for several aptitudes: "Scanning, across both auditory and visual sources, Perceptual Speed and Accuracy, Translating Information, Chunking, Interpreting Information, Sustained Attention, Recall from Interruption, Situational Awareness, Long-Term Memory, Problem Identification, Prioritization, Time-Sharing, Information Processing Flexibility, and Task Closure/Thoroughness. Two new aptitude requirements were identified: Dispositional Trust in Automation; and Computer-Human Interface (CHI) Navigation" (Broach, 2013, p i).

Several F-JAS surveys were conducted by Eißfeldt (2002) for three ATCO positions: tower, approach and area control. He found greater similarity between the job profiles of the two radar positions approach and area control in comparison to the tower position. Overall, the cognitive (e.g., speed of closure, visualization, selective attention, and time sharing), and interpersonal domains (e.g., cooperation, communication and stress resistance) received the highest ratings by SMEs. The same instrument was used in a recent study in Taiwan. Li (2019) found some differences in the importance ratings but could also confirm that time-sharing and stress resistance were rated as the most relevant abilities for Taiwanese ATCOs.

A recent comprehensive review of job analyses for ATCOs is provided by Wium & Eaglestone (2022).

5.4. Findings from Job Analyses for Pilots

A large number of job-analytic studies has addressed piloting tasks in military aircraft. They cover fixed wing pilots, rotary-wing pilots, and remotely-piloted vehicles. Fewer studies are available for civilian flight operations.

According to Mitchell & Driskill's (1996) review of military job analysis, Flanagan (1948) may have conducted some of the earliest job analyses for fixed-wing military pilots. His approach placed heavy emphasis on causes for training failures ("critical incidents"). However, the early CIT-method did not yet provide an adequate description of the KSAOs required in flight training. Throughout the 1960's, '70's and early '80's, new job analysis methods were developed and older ones refined (see Wilson, 2014 for examples). To identify which method was the best for aviation jobs, Driskill, Weissmuller, and Barrett (1989) compared 36 job analysis methods and identified Fleishman's Ability Requirement Scales as having the most comprehensive ability taxonomy and providing the best linkage to existing tests. They did note, however, that the scales did not adequately assess interpersonal and communication skills. Subsequently, Fleishman (1996) amended 21 social interpersonal scales and named the method Fleishman's Job Analysis Survey (F-JAS, see above).

Since the 80s and 90s the US Air Force periodically conducts task analyses on categories of aircraft (tanker, transport, etc.) to ensure that its pilot training is developing the skills necessary to perform the student's future flying assignments. To perform these analyses, it combines different methods including surveys, SME interviews, Fleishman's Ability Requirements Scale, and abilities and skills identified in O*Net. An example is an analysis of the KSAOs for fighter pilots in which Peña, Shore, Haight, Wolliston, and Gonzalez (2019) identified those KSAOs necessary for successful completion of training and for operation success. They identified instrument comprehension, achievement, situational awareness, aviation knowledge, adjustment, control precision, attention to detail, critical thinking, responsibility, and courage as the most important attributes for new fighter pilots.

Military helicopters are a different category of aircraft from fixed-wing aircraft and require their own job analyses. These job analyses tend to focus more on specific types of helicopters or missions than the corresponding task analyses for fixed-wing aircraft. For example, Jones and McAnulty (1984) compared the ability requirements for four different US Army helicopter missions and found no differences. The only ability they identified as important was general cognitive ability. In a more recent job analysis for US Army helicopter pilots, Kubisiak and Katz (2006) identified situational awareness, psychomotor ability, information processing, and decision making as the most important attributes.

The use of remotely piloted vehicles (RPVs) necessitated an understanding of the KSAOs required by RPVs and the development and validation of new selection systems (Howse, 2011; Phillips, Arnold, & Fatolitis, 2003; Portman, Biggerstaff, Blower, & Chapman, 1997). One of the earlier job analyses was conducted by (Barnes, Knapp, Tillman, Walters, & Velicki, 2000), who examined the need for rated aviators as pilots. The most highly rated attributes were oral comprehension, written comprehension, oral expression, and memorization. Mangos, Vincenzi, Shrader, Williams, and Arnold (2014) compared the required KSAOs across US navy and Marine Corps UAVs. The highest ranked attributes were communication, quantitative, problem solving/reasoning, and social skills.

Fewer task analyses are available for civilian aircraft and operations. Latarolla, Pliske, Hutton, and Chrenka (2001) reported a cognitive task analysis for weather flying behaviours of business pilots. The German Aerospace Center (DLR) conducted a number of surveys with an extended version of Fleishman's F-JAS (Goeters, Maschke and Eißfeldt, 2004; Maschke & Goeters, 2003). In a sample of 141 experienced airline pilots the interpersonal/interactive skills cooperation, communication, stress resistance, and decision making received the highest ratings. Equally high ratings were found in the cognitive domain for time sharing and spatial orientation. The lowest ratings were found for physical abilities.

5.5. Summary

The purpose of this chapter is to emphasize the importance of conducting a job analysis prior to administering personality or ability tests for personnel selection. The need to relate every single selection criterion to relevant job demands is part of the scientific standards of international organisations in applied psychology (APA, IPA). Additionally, some recent updates in the European aviation regulation require adjusting psychological assessments of pilots to the operational requirements within the specific organisation. As Broach et al. (2019) pointed out, the reality of many selection programs in worldwide aviation is that they do not have a strong scientific basis. Often, they are lacking relevant and up to date job information. The examples in the chapter have shown that there are plenty of job analysis studies published in the literature. Nevertheless, selection tools which are currently in use are still rarely linked to particular job requirement profiles. Therefore, with the review of tools and techniques in this chapter we want to put extra emphasis on the significance of job-analysis in selection.

5.6. References

Anderson, N. (1990). Repertory Grid Technique in employee selection. Personnel Review, 19(3), 9-15. https://doi.org/10.1108/00483489010143032 .

Barnes M.J., Knapp, B.G., Tillman, B.W., Walters, B.A. & Velicki, D. (2000). Crew systems analysis of Unmanned Aerial Vehicle (UAV) future job and tasking environments. Army Research Laboratory. ARL-TR-2081.

Broach, D. (2002). Air traffic controller ability requirements in the U.S. National Airspace System. In: H. Eißfeldt, M.C. Heil & D. Broach (Eds.), Staffing the ATM system. The selection of air traffic controllers (pp. 3 - 12). Aldershot: Ashgate.

Broach, D. (2013). Selection of the next generation of air traffic control specialists: Aptitude requirements for the air traffic control tower cab in 2018. Final report DOT/FAA/AM-13/5, Office of Aerospace Medicine Washington, DC.

Broach, D., Schroeder, D. & Gildea, K. (2019). Best practices in pilot selection. FAA Final report DOT/FAA/AM-19/6. Federal Aviation Administration, Office of Aerospace Medicine, Washington, DC 20591.

Christal, R. E. (1974). The United States Air Force occupational research project (Rep. No. AFHRL-TR-73-75). Lackland AFB, TX: Occupational Research Division (DTIC No. AD-774 574)

Crandall, G., Klein, G. & Hoffman, R.R. (2006). Workings minds: A practitioner's guide to Cognitive Task Analysis. Cambridge/MA: MIT Press.

Driskill, W. E., Koonce, J. M., Nance, P., & Weissmuller, J. J. (2001). Skills, Knowledge, and Abilities Underlying Success in USAF Pilot Training. (Technical Report No. AFRL-HE-AZ-TR-2000-0005). Mesa, AZ: Air Force Research Laboratory.

Eissfeldt, H. (1988). Beurteilungsrelevante Leistungskriterien im Rahmen der praktischen Ausbildung zum gehobenen Flugverkehrskontrolldienst -Eine Untersuchung anhand der Critical Incident Technique. DLR-IB-316-88-07.

Eißfeldt, H. (2002). ATC specific tests. In: H. Eißfeldt, M.C. Heil & D. Broach (Eds.), Staffing the ATM system. The selection of air traffic controllers (pp. 85 - 96). Aldershot: Ashgate.

EUROCONTROL (1997). Model of the cognitive aspects of air-traffic control. EATCHIP document number HUM.ET1.ST01.1000-REP-02. Brussels: EUROCONTROL.

EUROCONTROL (2006). Requirements for European class 3 medical certification of air traffic controllers. Brussels: EATM Infocentre, EUROCONTROL Headquarters.

European Union Aviation Safety Agency (2019). Acceptable means of compliance (AMC) and guidance material (GM) to Part-MED medical requirements for air crew. Issue 2, 28 January 2019. Retrieved from https://www.easa.europa.eu/document-library/acceptable-means-of-compliance-and-guidance-materials/amc-gm-part-med-issue-2 .

European Commission (EC) (2018). Commission regulation (EU) 2018/1042 of 23 July 2018. Official Journal of the European Union, L 188/3. 25.07.2018. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1539167289501&uri=CELEX:32018R1042.

Flanagan, J.C. (1948). The aviation psychology program in the army air forces. AAF Aviation Psychology Program Research Report No. I. Washington: U. S. Government Printing Office.

Flanagan, J.C. (1954). The critical incident technique. Psychological Bulletin, 51, 327–358.

Fleishman, E. A. (1975). Toward a taxonomy of human performance. American Psychologist, 30, 1127–1149.

Fleishman, E.A. (1996). Fleishman Job Analysis Survey (F-JAS-2): Social interpersonal scales. Management Research Institute. Potomac, MD.

Fleishman, E.A., Constanza, D.P. & Marshall-Mies, J. (1999). Abilities. In N. Peterson, M. Mumford, W. Borman, P. Jeanneret & E. Fleishman (Eds.), An occupational information system for the 21st century: The development of O*NET (pp. 175-195). Washington, DC: American Psychological Association.

Fleishman, E.A. & Quaintance, (1984). Taxonomies of human performance: The description of human tasks. Management Research Institute. Potomac, MD.

Fleishman, E.A. & Reilly, M.E. (1992). Administrator's guide: F-JAS. Palo Alto, CA: Consulting Psychologists Press.

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

Guilford, J.P. & Lacey, J.I. (1947). Printed classification tests. AAF Aviation Psychology Program Research Report No. 5. Washington: U. S. Government Printing Office.

Hansen, J.L. (1999). The recruitment and selection procedures in Denmark – "The right person". In EATMP (Ed.), Proceedings of the First EUROCONTROL Selection Seminar - "Current and required future selection work and methods in the ECAC area" (p. 123-134). EUROCONTROL.

Hermans, P. & Mulder, H.W. (1998). Job analysis and the selection interview. In K.M. Goeters. (Ed.) Aviation Psychology: A Science and a Profession (pp 81-95). Aldershot: Ashgate.

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

Jones, D.H. & McAnulty, D.M. (1984). An examination of ability requirements for various rotary wing missions. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 28(4), 361-365. https://doi.org/10.1177/154193128402800416

Knapp, D.J., Morath, R., Quartetti, D.A. and Ramos, R.A. (1998). A strategic approach to the analysis of air traffic controller jobs. Final report under FAA contract DTFA01-95-C-00052. Oklahoma City, OK: Civil Aeromedical Institute Training and Organizational Research Laboratory.

Kubisiak, C., & Katz, L. (2006). U.S. Army aviator job analysis (TR-1189). Arlington, VA.: U.S. Army Research Institute for the Behavioral & Social Sciences.

Latarolla, K., Pliske, R., Hutton, R. & Chrenka, J. (2001). Cognitive Task Analysis of Business Jet Pilots' Weather Flying Behaviors: Preliminary Results. Report NASA/TM-2001-211034. Langley Research Center, Hampton/VA.

Li, H.Y. (2019). Job analysis of air traffic controllers in ROC. Journal of Aviation Medicine and Science, 33, 13-26.

Mangos, Vincenzi, Shrader, Williams & Arnold (2014). UAS cross platform JTA. Final report. NAMRU-D report number 14-44. Naval Medical Research Unit Dayton/OH.

Maschke, P. & Goeters, K.-M. (20030). Ab-initio flight training and airline pilot performance: Differences in job requirements. In G. Edkins & P. Pfister, Innovation and Consolidation in Aviation (pp189-

195). Ashgate, Aldershot.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 56, 347–368.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1989). Position Analysis Questionnaire. Logan, UT: PAQ Services, Inc.

Militello, L.G. & Hutton, R.B. (1998). Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. Ergonomics, 41(11), 1618.1641.

Mitchell, J.L. & Driskillm W.E. (1996). Military job analysis: A historical perspective. Military Psychology, 8(3), 119-142. DOI: 10.1207/s15327876mp0803_1

Morgeson, F.P., Brannick, M.T. & Levine, E.L. (2020). Job and work analysis: Methods, research, and applications for human resource management (3rd ed.). Los Angeles/CA: Sage Publishing.

Peña, D.A., Shore, W.C., Haight, N.R., Wolliston, D.J., & Gonzalez, M. (2019). USAF assessment of needs analysis: 11F fighter pilot. San Antonio, TX: Operational Technologies Group.

Peterson, N.G., Mumford, M.D., Borman, W.C., Jeanneret, P.R., Fleishman, E.A. (1999). An Occupational Information System for the 21st Century: The Development of O*NET. American Psychological Association, Washington D.C.

Philipps, H.L, Arnold, R.D. & Fatolitis, P. (2003). Validation of an unmanned aerial vehicle operator selection system. Proceedings of the 45th Annual Conference of the International Military Testing Association, 129-139. Pensacola/FL, 3-6 November 2003.

Portman, C., Biggerstaff, S., Blower, D. & Chapman, A. (1997). Unmanned Aerial Vehicles (UAV) Pilot Selection System. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 41(2), 1385–1385. doi:10.1177/107118139704100217 7 .

Reeb, C. & Gabauer, A. (2016). A Comparative Task Analysis for Military Air Traffic Controllers and Air Battle Managers. In M. Schwarz & J. Harfman (Eds.) Proceedings of the 32nd EAAP Conference, Cascais, Portugal, 397-405

Seamster, T.L., Redding, R.E. & Kaempf, G.F. (2017). Applied cognitive task analysis in aviation. London: Routledge.

Taylor, P.J., Li, W.D., Shi, K. & Borman, W. (2008). The transportability of job information across countries. Personnel Psychology, 61, 69-111.

Wilson, M.A. (2014). A history of job analysis. In L.L. Koppes (Eds.) Historical Perspectives in Industrial and Organizational Psychology, pp. 219–41. Mahwah, NJ: Erlbaum.

Wium, J., & Eaglestone, J. (2022). Job analysis for air traffic control: A review of the literature from 1949-2020. Aviation Psychology and Applied Human Factors. 12, pp. 31.49. https://doi.org/10.1027/2192-0923/a000218.

6. Selection Methods & Instruments

Jóhann Wium, Jennifer Eaglestone, Karien Stadler & Hans-Jürgen Hörmann

In order to identify candidate personal attributes that are required for a job (as determined by a job analysis), a candidate is assessed. An assessment is any form of systematic information gathering pertaining to these attributes that is then used to draw inferences about the candidate (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). It represents the broadest term used when determining candidate characteristics.

Assessments can be performed in a variety of different ways, depending on the means of information gathering or the psychological construct being assessed. A selection method refers to a grouping of similar assessments that are meaningfully connected, representing a specific subset of assessments. Sometimes a selection method will include assessments that simply vary in their specific content (e.g. knowledge tests), while other methods can contain diverse assessments (e.g. assessment centers). This chapter covers a variety of common selection methods, how they have been used in aviation and discusses their effectiveness or validity as selection methods.

When an assessment uses a predefined protocol or technique to determine applicant attributes, we refer to it as a selection instrument or a selection tool. For example, ability testing is a selection method while an established test of spatial reasoning represents a selection instrument. Similarly, an interview is a selection method while a structured biographical interview is a selection instrument.

In 1998, Schmidt and Hunter published a now well-known meta-analysis on personnel selection methods, to provide an estimate of the generalized validity of each selection method. Since then, a number of other meta-analyses have looked both at specific selection methods and the selection of specific professions. The meta-analysis methodology has also been refined, with some research suggesting that traditional meta-analyses under-estimate the validity of selection methods (Schmidt et al., 2016), while others argue that they over-estimate it (Sackett et al., 2021).

While it is true that meta-analyses are not perfect indicators of validity, they do condense a large amount of data into clear and concise metrics and therefore provide a useful starting point when comparing disparate and different methods in selection. Traditionally, meta-analyses provide two distinct estimates of generalized validity: a calculated indicator derived directly from statistical information of other studies (uncorrected r); and a statistically corrected indicator, that is believed to give a better estimate of the effect being measured (corrected r). As meta-analysis methodology for correction can vary between authors, uncorrected generalized validities will be provided unless otherwise stated.

Where available, meta-analyses done on aviation professions (pilots and air traffic controllers) will be referenced when discussing selection methods. Rarely, there is no published research on a particular selection method for aviation personnel and general references are instead used. For more details on available meta-analyses done on aviation personnel, see chapter 8.

As discussed in chapter 5, a job analysis determines the knowledge, skills, abilities and other characteristics required for a job (KSAO's). This chapter follows this same structure and discusses selection methods that are often used to assess knowledge, skills, abilities and other characteristics. It is important to note, however, that while specific methods are discussed under specific headings, a method can at times be used to assess multiple different attributes. For example, while interviews are discussed under the heading of "Other characteristics", interviews can also be used to determine skills or knowledge.

Keep in mind that while we use results from meta-analyses to compare validities of different methods, practitioners should consider the exact validity of a specific selection instrument when deciding which selection instruments to use. A discussion of some of the criteria selection instruments should be judged on is included at the end of this chapter. How selection methods and selection instruments are combined into a single process that comprehensively assesses an individual for a role is then described in chapter 7.

6.1. Assessing Knowledge

6.1.1. Knowledge Tests

Definition: Knowledge tests are frequently used in pilot and ATCO selection and are measures of prior knowledge that has usually been obtained through formal education, such as knowledge of physics, mechanical comprehension, English, and mathematics. In some cases, it can be difficult to distinguish between an ability test, e.g. a test that assesses quantitative ability, or a test that assesses acquired knowledge, for instance, mathematics. A helpful rule of thumb would be to consider how the test is scored, if the test looks at simple and absolute percentages of right vs. wrong answers then it is more likely to be a knowledge test but if the test looks at the scores in comparison with a norm group, then it is more likely to be an ability test.

Pilots or potential air traffic controllers are sometimes tested on their knowledge of aviation-related information or subjects. This can, however, become problematic if selecting from only ab initio candidates that have had no formal education in aviation-related subjects, particularly if the applicant pool is mixed with some candidates having aviation-related experiences and others not. In some cases, this is done intentionally, as candidates with previous aviation-related experience or knowledge are considered to be more likely to succeed in the role than others.

Use in aviation: Knowledge tests have effectively been in continuous use in the US Air Force pilot selection battery since World War II (Carretta & Ree, 1995). They are also in use in other selection batteries, such as the DLR battery (Zierke, 2014) and the Norwegian Air Force's battery (Martinussen & Torjussen, 1998). In civilian pilot selection, aviation knowledge or aviation information tests are common and have been in use since the mid 1930's (Viteles, 1945).

Similarly, general and aviation knowledge tests have been used since the very beginning of ATCO testing in 1956 (Trites, 1961) and have remained a staple in ATCO selection since then. For a specific example of the development of an aviation information test for ab initio air traffic controllers, see Dailey & Pickrel (1984).

Evidence for validity: The meta-analysis performed by Martinussen (1996) shows aviation information tests to be good predictors of pilot performance ($r_{uncor} = 0.22$) with Hunter and Burke (1994) finding a very similar effect ($r_{uncor} = 0.22$). The efficacy of other knowledge tests in aviation is more uncertain. Martinussen (1996) grouped together other knowledge tests and academic grades, and when taken together they had a generalized validity of $r_{uncor} = 0.15$. Hunter and Burke (1994), however, found that tests of general information had an $r_{uncor} = 0.25$, but did not define exactly what assessments were included in their general information grouping.

For air traffic controllers, Martinussen, Jenssen and Joner (2000) found the generalized validity of occupational knowledge to be $r_{uncor} = 0.18$. Occupational knowledge is unfortunately not specifically defined so it is uncertain whether it refers to aviation information tests, general knowledge tests, both or some other assessments[5] .

Note that there is mixed evidence regarding whether knowledge tests provide incremental validity beyond measures of general intelligence. While some have found that specific knowledge tests add little (Olea & Ree, 1994), others found that it added incremental validity (Zierke, 2011) and might be assessing different constructs than general ability tests, e.g. motivation to learn (Zierke, 2014).

Further reading: For the results of meta-analyses of the efficacy of knowledge tests outside of aviation, see Dye, Reck and McDaniel (1993).

[5] It is probable that „occupational knowledge" is a grouping variable for the results from the Occupational Knowledge Test (Dailey & Pickrel, 1984) and other similar measures of aviation-information, but it is not specifically stated that this is the case.

6.2. Assessing Skills

6.2.1. Work Sample Tests

Definition: The premise of work sample tests is that the best predictor of future behaviour is how you perform now in the same or similar situation. Work sample tests are therefore tests where the applicant has to perform tasks that are similar to those performed on the job.

Work sample tests are generally divided into two categories, low and high fidelity. Low-fidelity work samples are simplified versions of the job or task at hand, whereas high-fidelity work samples are more similar to the job itself. For experienced pilots and ATCOs, high-fidelity work sample tests are usually administered in high-fidelity simulators. Because ab initio candidates are assumed not to have established job skills they are oftentimes rather assessed in low-fidelity simulators (such as on a PC) where the salient aspects of the exercise can be controlled to create a simple task environment.

Like other selection instruments, a simulator exercise must be standardized and evaluated either using an automated scoring system or highly trained raters. Simply letting a candidate do a run in a simulator would therefore not count as a work-sample test.

Use in aviation: Simulators have been used for pilot selection since the late 1930's (Signori, 1949; Viteles, 1945) and research has been conducted to identify the most sensitive and reliable performance measures (Connelly & Shipley, 1982) and maneuvers (Spinner, 1989). Because ab initio candidates generally do not know how to fly, psychologists have standardized initial training packages that are administered before the candidate is evaluated (Davis, Koonce, Herold, Fedor, & Parsons, 1997; Spinner, 1991; Woychesin, 2002). Some of these ab initio selection systems may use speed of learning measures as well as performance measures as predictive variables (Davis et al., 1997). For specific examples of flight simulator use in pilot selection see Meierfrankenfeld, Gress & Vorbach (2015), Bramble & Koonce (1998) or Fowler (1981).

Whenever possible work-sample tests should be included in selection for ATCOs, ideally combining the processing of acoustic and visual information simultaneously (Eißfeldt, 2002). Not only do work sample tests allow for valid assessment of the many relevant task abilities but they are also extremely motivating for those interested in the job. There are however certain aspects that must be controlled for, including prior experience with gaming (although this might change with time) and commercially available coaching. For more information on commercial coaching see chapter 10. For examples of ATC specific work-sample tests and further references see Eißfeldt (2002) or Hättig (1991).

Evidence for validity: For pilots the generalized validity of work sample exercises has been found to be runcor = .34 (Hunter & Burke, 1994) and r_{uncor} = .28 for air traffic controllers (Martinussen, Jenssen & Joner, 2000). This is not dissimilar to results found by Roth, Bobko and McFarland (2005) outside of aviation where they found it to be r_{uncor} = .26. In Hunter & Schmidt's original article, they presented work sample tests as one of the best single predictors of future job performance, but later research (e.g. Schmidt et al., 2016) found them to be less predictive than previously stated.

Further reading: For more on the specific predictive validity of simulators for pilots, see Koonce (1979). For meta-analyses done on the efficacy of flight simulators for training, see Hays et al. (1992) and de Winter, Dodou and Mulder (2012). For more general information on the use of high and low fidelity work samples in high stakes selection, see Lievens & Patterson (2011).

6.2.2. Situational Judgement Tests

Definition: Situational Judgement Tests (SJTs) can be categorized as low-fidelity job-samples because they present applicants with job-related situations with possible responses to those situations. Applicants must indicate which response they would choose that can then be used to assess their judgement and decision-making capabilities or job-related knowledge (Lievens et al. 2008).

Use in aviation: There is little published research available at this time that describes the development or validity of SJTs for operational personnel in aviation. In a recent study by Goerke and Maier (2022), a teamwork SJT showed incremental validity over cognitive tests and a personality inventory for the prediction of assessment center results in a sample of ab-initio pilot applicants and some agencies have begun using SJTs for other roles in aviation, with one such example the inclusion of a SJT for promotions in the U.S Air Force (Barron et al., 2021; Sullivan et al., 2019).

Evidence for validity: Outside of aviation, a large-scale study done by Weekley and Jones (1999) on almost 4,000 employees found that an SJT could predict supervisor ratings ($r = 0.19$) and was connected to, but distinct from, general mental ability and experience. This is supported by the results from the meta-analysis done by McDaniel et al. (2007) that SJTs have a generalized validity of r_{uncor} = 0.2.

Further reading: More information on use of SJTs in selection can be found for instance in McDaniel et al. (2001), Corstjens, Lievens, & Krumm (2017), McDaniel & Nguyen (2003), or Whetzel and McDaniel (2009). For meta-analysis specifically on the constructs assessed by SJTs, see Christian, Edwards & Bradley-Geist (2010).

6.3. Assessing Abilities

6.3.1. Cognitive Ability Tests

Definition: An exact definition of cognitive abilities has remained elusive for psychologists and while cognitive abilities refer to any ability connected to an internal mental process they can come in a vast variety of different forms. Cognitive ability can, for example, be in the form of verbal fluency, numerical aptitude, spatial comprehension, executive function, abstract reasoning, problem solving and many, many more. Oftentimes the definition of cognitive ability has been linked to intelligence and there are multiple theories that attempt to conceptualize and explain it.

For practitioners this can mean that a single selection process might use multiple different tests based on different theories of intelligence. In and of itself this is not a problem as long as each test is validated. A cohesive framework for intelligence, however, is necessary when looking into ways to improve a selection procedure or for test development. For a comprehensive theory of intelligence for use in aviation we recommend the Cattell-Horn-Carroll theory of intelligence and its three-level hierarchy of general ability (g), broad abilities (ten domains of intellectual abilities) and narrow abilities (specific cognitive abilities) (see Carroll, 1993; Ackerman & Heggestad, 1997; Taub & McGrew, 2004).

As a selection method, cognitive ability tests refer to a structured set of items with specific right and wrong answers and are sometimes referred to as tests of maximum performance (The British Psychological Society, 2017).

Use in aviation: Similar to knowledge tests, cognitive tests have a long history of use in aviation. Cognitive ability tests are common parts of selection batteries and have been the subject of considerable research, recommendations and guidelines of use.

In one such guideline, IATA (2019) distinguishes between "basic mental abilities", "composite mental abilities" and "operational abilities" in their guidance material for psychological testing. While IATA provides examples and states that these abilities connect directly to pilot competence (e.g. with composite mental abilities connecting to automation, manual control and aeroplane flight path management) they do not define specifically what constitutes a basic mental ability, composite mental ability or operational ability.

EASA's Accepted Means of Compliance for CAT.GEN.MPA.175 (EASA, 2022) stipulates that a psychological assessment of flight crew include an assessment of cognitive ability. While the regulation does not explicitly use IATA's distinctions, they go on to suggest its use in the regulation's published guidance material. It is therefore possible that practitioners will be asked how their selection tests connect to IATA's framework.

At this time no published research supports the distinction between basic, composite and operational cognitive abilities. While it is possible to use that distinction when grouping attributes as part of a pilot's job requirement profile (see chapter 5 and 7) they are not specific psychological constructs that can be assessed as such.

Evidence for validity: Cognitive ability tests have consistently been found to be one of the best predictors of general job performance with selection batteries often including several tests of cognitive ability. Specifically in aviation, meta-analyses have shown that the generalized validity for cognitive ability tests was $r_{uncor} = 0.13$ for pilots, and $r_{uncor} = 0.21$ for air traffic controllers. The range in efficacy for specific (i.e narrow) cognitive abilities for pilots is between $r_{uncor} = 0.12 – 0.29$ and between $r_{uncor} = 0.19 – 0.25$ for air traffic controllers (Hunter & Burke, 1994: Martinussen, 1996, Martinussen, Jenssen & Joner, 2000).

Further reading: For a more general overview on cognitive ability tests see for instance Cohen, Schneider and Tobin's 'Psychological Testing and Assessment: An Introduction to Tests and Measurement' (2022) but for professional guidelines on the use of tests see the International Test Commission guidelines (International Test Commission, 2013). For more information on how to utilize cognitive ability tests in selection of operational aviation personnel see IATA's 'Best Practices for Aptitude Testing' (IATA, 2019) or EUROCONTROL's 'Selection Tests, Interviews and Assessment Centres for Ab Initio Trainee Controllers: Guidelines for Implementation' (EATMP, 2002). For a review of how ability testing has been used in civilian and military pilot selection see Damos (2011), Paullin et al. (2006) and Turnball (1992), for a comparison of pilot aptitude test batteries see Broach, Schroeder and Gildea (2019) and Damos (1996).

6.3.2. Psychomotor Abilities

Definition: Ackerman (1988) defined "A general psychomotor ability represents individual differences predominantly in the speed of responses to test items with little or no cognitive processing demands" (p 290). While test development for cognitive abilities aimed at tasks representing a single cognitive function, tests of psychomotor abilities are often more complex with respect to the abilities sampled. Psychomotor tests can range from simple reaction time or tapping tests, via maze tracing tasks and peg boards, to tracking tasks, complex apparatus tests and simulation tests. Classical apparatus tests include for example Rotary Pursuit, Two-hand Coordination, Complex Coordination with hand and foot control tasks or the Rudder Control Test for Motor Kinesthesis (see Fleishman, 1953 or Melton, 1947). The sort of information processed in these tasks stems primarily from visual, auditory, tactile and vestibular stimuli in combination with sensory feedback resulting from motor responses.

Extensive dimensional analyses of psychomotor abilities were conducted historically by Fleishman (1953, 1964) and later recapped by Carretta and Ree (1997) and Chaiken, Kyllonen and Tirre (2000). However, the particular number and nature of several independent factors underlying psychomotor performance evolved with the development of new test principles and also with the amount of practice on these tasks (e.g., Fleishman & Hempel, 1954). In order to verify postulated subcomponents of psychomotor performance, Chaiken et al. (2000) analysed sixteen psychomotor tests with confirmatory factor analyses. The sixteen psychomotor tests were supposed to represented the four most general factors of Fleishman's taxonomy (Chaiken et al., p 203-207):

- Control Precision: Tests involving fine arm-hand or leg movements that require precise and quick positioning in the operation of equipment (e.g., Rotary Pursuit)

- Multilimb Coordination: Tests involving simultaneous coordination of two of more limbs when operating devices with several controls (e.g., Complex-Coordination Tests, Two-hand Coordination)

- Rate Control: Tests involving continuous corrections of motor responses in synchrony with variably speeded objects (e.g., Lane Tracking, Wheel Avoidance)

- Response Orientation: Tests involving rapid choice of correct motor responses in correspondence to discrete stimuli (e.g., Choice Reaction Test)

The authors found that a model with a single general factor resulted in a better data fit than a model with four Fleishman-specific psychomotor factors. A model with one general psychomotor factor plus four nested lower order factors was only marginally better. Therefore, a taxonomic representation of different psychomotor abilities does not seem necessary.

Use in aviation: Tests of psychomotor abilities are among the most common selection methods especially for ab-initio pilots. Compared to most cognitive ability tests, there is more extensive test equipment for displaying dynamic stimuli and recording discrete of continuous motor responses. In the past, complex apparatus tests were utilized to assess psychomotor abilities of the candidates. Melton (1947) provided an overview of these vintage tests. An update is available by Griffin and Koonce (1996). Nowadays, the power of regular desktop PCs is sufficient to present various psychomotor tasks and to record the applicant's performance. A typical test paradigm used especially for ab-initio pilots is that of manual tracking tasks. As explained in section 3.3, a tracking task involves a visual target stimulus (e.g., a small cross or a circle), which has to be approximated with the cursor for a certain time. The cursor is controlled by the applicant usually with a control stick. An overlaying disturbance function can influence the behaviour of the target and the cursor. The averaged or accumulated deviation (absolute or squared) between the target and the cursor is a reciprocal measure of psychomotor performance. Tracking tasks can be combined with further simultaneous tasks to represent a multi-tasking scenario. Typically, the tracking task is regarded as the primary task and any parallel task (e.g., an acoustic monitoring or a simple arithmetic task) as secondary.

With the highest degree of complexity, motor coordination can also be assessed in a (flight-) simulator. As a selection test for licensed pilots, a realistic flight simulator is counted as a work sample test (see section 6.2.1). Performance in a work sample test is typically determined by multiple factors (e.g., level of experience, knowledge, motor coordination, spatial orientation etc). Simulators can be fixed-based or with motion.

Genuine psychomotor abilities are usually not among the most relevant job requirements for ATCOs. However, in some cases psychomotor tests are applied as part of a multi-tasking test environment or simply as measures of attention. For example, the test of multi-tasking abilities of the FEAST battery for European ATCOs includes psychomotor tasks to address reaction time, concentration in addition to multi-tasking (Damitz, Chetcuti & Henriques, 2010).

Evidence for validity: Wheeler and Ree's (1997) provided a validation study of general and specific psychomotor tracking tests. The general psychomotor ability score was the best predictor for the pass/fail training criterion of N = 1099 US Air Force pilots (range-restriction corrected r = .19). Also, performance in six flying tasks could be predicted significantly with the general psychomotor score in a subsample of N = 833 pilots (range-restriction corrected r = .28). Only reaction time tests provided some incremental validity. However, reaction time is often linked to cognitive abilities rather than to psychomotor abilities (e.g., Jensen, 1982).

Tests of psychomotor abilities have consistently proven to be among the best predictors of flight training success of pilots. The meta-studies of Hunter and Burke (1995) and AlMamari and Traynor (2019) found average correlations ranging from r = .18 to r = .49. These scores are only matched by flight simulator tests, which have average predictive validities ranging from r = .19 to r = .55 in Hunter and Burke (1995) and r = .24 to r = .35 in AlMamari and Traynor (2019). For ATCO selection, psychomotor tests seem to be less predictive. Chapter 8 provides more details of validation efforts for pilot and air-traffic controller selection.

Further reading: Recommended further reading especially includes papers on the concept of psychomotor abilities (Ackerman, 1988; Fleishman, 1953), their structure and sub-components (Fleishman, 1964; Chaiken, Kyllonen & Tirre, 2000) as well as their application in a multi-tasking test environment (Damos, 2020).

6.4. Assessing Other Characteristics

6.4.1. Personality Tests

Definition: As with cognitive ability, an exact definition of personality remains elusive. It can be considered to be an internal factor that has some causal effect on behavior but in addition to personality there are also various assessments to determine a candidate's interest, motivation, cognitive style or any other behavioral preferences. For our purposes, any assessment that uses a structured set of questions designed to determine an individual's preferential behavior is considered a personality test. These tests are sometimes also referred to as tests of typical performance (The British Psychological Society, 2017).

Use in aviation: Similar to cognitive tests, EASA's AMC to CAT.GEN.MPA.175 requires an assessment of personality traits of flight crew. Interest in selecting pilots with the "right" personality is not new, dating back to World War I (Rippon & Manuel, 1918) and has continued to the present day (Carretta et al., 2014; Dolgin & Gibb, 1988; Dolgin & Gibb, 1989; Hilton & Dolgin, 1991; Hörmann, Radke, & Hoeft, 2007; Maschke, 2004). However, little progress connecting personality to job performance was made until the development of and widespread acceptance of the "Big Five" personality dimensions (McCrae & Costa, 1989, 1997). Subsequently, personality testing has continued using predominately Big Five selection instruments. Despite this the overall relationship between personality and job performance has consistently been shown to be lower than cognitive ability tests.

Evidence for validity: Meta-analyses done on aviation personnel show that generalized validity for personality is between r_{uncor} = 0.1 – 0.13 for pilots and r_{uncor} = 0.03 for air traffic controllers (Hunter & Burke, 1994: Martinussen, 1996, Martinussen, Jenssen & Joner, 2000).

Outside of aviation, research has used a stronger connection between job performance and personality (e.g. for conscientiousness; Schmidt & Hunter, 1998) but it is possible that the difference stems from the objective versus subjective nature of performance in different roles, with generalized validity for objective performance standards (r_{cor} = 0.12) being much closer to those found for roles in aviation (see Tett et al, 1991).

The utility of personality assessments in selection for aviation roles is therefore still debated and transparency and reliability are major issues affecting the predictive validity of those instruments (Maschke, 2004). As an alternative to personality, some have decided to focus rather on social competence rather than behavioral preference (see Hörmann & Goerke (2014) and Hörmann, Radke & Höft (2007) for more details).

Further reading: For a critical overview of personality instruments see for instance Goldstein, Beers and Herse (2004). For a review on the use of personality assessment for pilot selection see Dolgin & Gibb (1988), Maschke (2004) and Damos (2011). For more on the use of personality assessment for air traffic controllers see Dean et al. (2002).

6.4.2. Biographical Data

Definition: Biographical data (biodata) is often collected in the form of a biographical questionnaire, with questions on factual information such as basic demographics, education details (including grades), previous employment, family, interests, job motivation and more. This low-cost instrument is often used at the beginning of a selection procedure to decide which candidates will be invited to later rounds of selection.

Use in aviation: Biographical data has been used in aviation since the beginning of pilot selection (see chapter 3) and still used today. While more frequently used for ab initio military pilot selection (see e.g. Stricker, 2005) than ab initio civilian pilot selection it is generally used in the earliest stages of selection as they are inexpensive to administer for a large group of applicants and require a short testing period (Stokes et al., 1994).

Evidence for validity: Martinussen (1996) in her meta-analysis of psychological measures as predictors of pilot performance shows an r_{uncor} = 0.21 for biographical questionnaires. This is lower than the predictive value for biographical data of r_{uncor} = .35 reported by Rothstein et al. (1990). It has been theorized though (Schmidt & Hunter, 1998) that there is an overlap with tests of general cognitive ability, suggesting that there might be less value added by the biodata for professions which use extensive cognitive ability testing.

Meta-analysis has not been done on biodata for ATCO selection, but Collins et al. (1990) found the predictive power of the FAA's BQ biodata inventory added to the predictive power when combined with cognitive ability measures, increasing the validity coefficient from .41 (just cognitive ability) to .48 (cognitive ability plus biodata). More recently a study by Broach (2012) yielded similar findings with biographical data accounting for an additional 2% of the variance above the 27% of variance accounted for by the cognitive ability tests. Although small, the authors concluded that this small increment in validity could be very useful considering the high number of applicants to select from and high training costs of those selected.

One should be careful in the use of biographical data. Eißfeldt (2004) forewarns of the possible adverse impact of such scales and emphasizes the importance of candidate privacy and continuous validation. The validity of this kind of instrument can change with time as applicant pools and jobs change and biographical questionnaires are susceptible to socially desirable answering, therefore requiring constant control for that factor.

Further reading: For more general information on the development, validation, and use of biographical data in selection see Stokes et al. (1994) and Breaugh (2009). For information on the development and use of biographical data in aviation see Eißfeldt (2004), Dean et al. (2002), Maschke (2004) and Stricker (2005).

6.4.3. Interviews

Definition: The method most frequently used to select employees is the employment interview (Dipboye, 1994; Schuler, Frier, & Kauffmann, 1993). Selection interviews can be divided into three different categories: structured, semi- structured and unstructured. The structured interview is an interview by which all questions are predetermined while a semi-structured interview has predetermined subjects and often main questions but the interviewer is free to decide on follow-up questioning. Structured interviews, such as the behavior description interview (Janz, 1982) and the situational interview (Latham, Saari, Pursell, & Campion, 1980) have yielded high criterion-related validity coefficients for a wide range of job positions, performance criteria, and demographic groups (Huffcutt et al., 2004; Latham & Sue-Chan, 1999; Taylor & Small, 2002).

The unstructured interview is a more free-flowing interview. One major drawback to the unstructured interview is the issue of fairness. Because different candidates may receive different questions, some candidates may believe that they have been treated unfairly. Use of an unstructured interview also opens the door to conscious or unconscious bias and, in some countries, increases the company's legal exposure. Using job requirements as the foundation for interview questions usually decreases legal exposure. Consequently, structured and semi-structured interviews are preferred when fairness, bias or legal requirements may be an issue.

An interview is, however, only as good as the interviewer. Fair and reliable interviews require well trained interviewers who are knowledgeable about the job. In addition, descriptively anchored rating scales are important and can contribute to substantial improvements in both rating accuracy and interrater reliability (Melchers, et.al., 2011).

Interviews can be used to assess a variety of attributes but they are commonly used to assess "soft skills" or "behavioural competencies", such as personality and interpersonal skills, and can also be useful to confirm biographical information.

Use in aviation: Interviews are a popular selection method in aviation. Suarez et al. (1994) surveyed corporate operators, regional airlines, as well as specialized and commuter air services and reported that the most common selection method across all carriers were interviews, which represented 96% of the responses. Similarly, EUROCONTROL (EATCHIP, 1996) asked air navigation service providers on selection methods and found that 83% used interviews and that the majority of interviews were intended to gauge motivation, personality and communication skills.

Despite its ubiquity there is not much published research on its specific use or efficacy in aviation. Some research and materials have been published on the use of interviews for selecting air traffic controllers and EUROCONTROL published guidelines for structuring interviews and distinguished between two interview types that could be used in that regard; biographical or situational interviews (EATMP, 2002).

A situational interview is a structured interview that asks questions what an applicant would do in a hypothetical situation that might arise at work. Its use for selecting student ATCOs was originally developed by the Swedish Civil Aviation Authority in the late 90s (Brehmer, 1995; Brehmer, 1998). A biographical interview focuses rather on an individual's past experience, achievements, motivation and similar. A meta-analysis done by Salgado and Moscoso (2002) supports that there are two distinct interview types and that they do appear to assess distinct attributes of the candidate.

Evidence for validity: As mentioned earlier, there is not much research examining the specific effectiveness of interviews in aviation. What research there is on the contribution of interviews in aviation has been mixed. Conzelmann & Keye (2014) showed that data from semi-structured interviews was useful as a predictor for performance and success in ATCO training but in one of the earliest studies of US ab initio civilian pilot selection, Viteles (1945) found too little incremental validity for a structured interview to justify its time and cost. The most recent study of US military pilot selection, Walters, Miller, and Ree (1993) demonstrated that a structured interview provided no incremental validity to the existing pilot selection battery. The structured interview was predictive when considered alone but appeared to be assessing attributes that were already assessed by the aptitude and personality tests in the battery.

In general, research has been fairly conclusive in showing that structuring interviews around the characteristics to be evaluated leads to a more effective interview process, as long as interviewers are well-trained and that candidate performance is evaluated using a well-developed rating scale. Outside of aviation, McDaniels et al. (1994) performed a meta-analysis on different interview types, purpose and content of general employment interviews and found structured interviews had somewhat better prediction (r_{uncor} = 0.28) than unstructured (r_{uncor}= 0.21). In a more recent analysis, however, Huffcutt et al. (2014) found that structured interviews outperformed unstructed interviews by a considerable margin (r_{uncor} = .35 for structured, vs r_{uncor} = 0.12 for unstructured).

Further reading: For an in-depth analysis of interview structure on reliability and validity, see Campion, Palmer, & Campion, 1997). For more on the different constructs that employment interviews can assess, see Huffcutt et al. (2001).

6.4.4. Assessment Centres

Definition: The International Taskforce on Assessment Centre Guidelines (2015) defines an assessment centre as “A process employing multiple assessment components, multiple assessors, and the use of simulation exercises to produce judgments regarding the extent to which an assessee displays proficiency on selected behavioral constructs” (p.1269). In general, the focus of assessment centres is on measuring interactive and interpersonal behavior and they can include both group and individual exercises, such as leaderless group discussion, individual presentations, and role-plays.

Use in aviation: The term "assessment centre" is sometimes used to describe a full day of testing, including other measures such as psychometric tests, situational judgement tests and interviews. However, in aviation the term 'assessment centre' is most commonly used as a combination of group and individual exercises with multiple assessors and multiple job-relevant criteria in job-representative scenarios.

Evidence for validity: Outside of aviation, meta-analyses done on the efficacy of assessment centers have found them to be effective at predicting supervisory performance ratings, ranging from $r_{uncor} = 0.17$ (Hermelin, Lievens & Robertson, 2007) to $r_{uncor} = 0.25$ (Gaugler et al., 1987). There has also been research into looking at the efficacy of assessment centers to predict specific "soft skills", ranging from $r_{cor} = 0.25$ for "Awareness of Others" to $r_{cor} = 0.39$ for "Problem Solving" (Arthur et al., 2003). However, as the quality of assessment centres (i.e. the combination of design and how it is used) tends to vary more than other selection methods the specific validities may also vary considerably (Thornton & Rupp, 2006). In aviation, assessment centres have been shown to have added value in the selection of both ATCOs (e.g. Höft and Pecena 2004) and pilots (e.g. Damitz, 2003).

Further reading: The Guidelines and Ethical Considerations for Assessment Center Operations (International Task Force on Assessment Center Guidelines, 2009) provide recommendations to those wishing to design and conduct an assessment centre. As with ability and personality testing, some countries also have national standards for the delivery and design of assessment centres (e.g. BPS, 2015; Swiss Assessment, 2004; Arbeitskreis Assessment Center, 2004). These can also be valuable resources for the design, delivery and validation of your assessment centre.

For the validity of common assessment center exercises, see Hoffman et al. (2015).

6.5. Instrument Criteria

Each selection method contains a multitude of specific selection instruments to choose from. It is the responsibility of the practitioners who decide which selection instruments to use, to use appropriate and effective instruments. In that regard, selection instruments must be reliable and valid, and the data supporting its reliability and validity must be based on a reasonable size sample. Detailed information on issues concerning the reliability and validity of selection instruments may be found in the Standards for Educational and Psychological Testing published by the American Educational Research Association (2014).

Test providers must be able to present reliability and validity information. However, practitioners should search the literature to find articles in refereed journals written by neutral, third parties that present data on a test's reliability and validity.

If no such data is available, practitioners should consult data on comparable tests or consider performing an internal study to determine concurrent validity. Local validation (within your own organisation) should ideally be conducted to determine the instrument's reliability and validity for the specific population to be assessed. See Rathje (2002) and Zinn et al. (2020) for overviews of criteria and information for evaluation of tests used in an aviation context.

Test batteries should be evaluated in the same way as an individual test: reliability and validity data should be available and meet acceptable levels. Studies in referred journals should be consulted. Because batteries cost more than individual tests and require more time to administer, the system developer should ensure that the battery assesses KSAOs that have been identified by the job analysis. See Appendix A Selection Checklist for a checklist of technical aspects to consider when using psychological tests.

Practitioners also needs to think critically about publisher-provided test norms. Many vendors base their norms either on the general population or on aviation personnel that are not representative of the population to be tested. Small differences in population characteristics can make large differences in the norms and can affect a test's utility in practice.

In addition to this there are also secondary requirements that the practitioner needs to consider. The usefulness of a test can also be influenced by factors such as ease of administration, cost, fairness, cultural robustness and many more. While important, many of these factors are unfortunately outside the scope of this chapter but for more information on some of the common aviation-specific factors that can affect the choice of selection instruments, see chapters 7, 8 and 10.

6.6. Summary

In order to ensure that aviation personnel are selected for the right attributes, selection instruments are used to assess potential candidates. With a wide variety of methods and instruments available it is important that practitioners are aware of the difference between selection methods and their respective instruments. Many of these instruments have been used in aviation for decades and have a wealth of information available on them (see chapter 3 and 4), while others can be so new that there is little if any established research on their efficiency (see chapter 11). This chapter is intended to provide a useful starting point when developing selection for aviation personnel but it remains the practitioner's responsibility to examine the evidence for the validity, reliability and usefulness of their selection instruments before putting them into practice.

6.7. References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. Journal of Experimental Psychology: General, 117, 288-318.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Arbeitskreis Assessment Center (2004). German Standards for Assessment Center Operations. Retrieved from: https://www.forum-assessment.de/images/standards/149_german-ac-standards-2004.pdf

Arthur, W., Day, E., McNelly, T., & Edens, P. (2006). A Meta-Analysis of the Criterion-Related Validity of Assessment Center Dimensions. Personnel Psychology. 56. 125 - 153. DOI: 10.1111/j.1744-6570.2003.tb00146.x.

Barron, L., Aguilar, I., Rose, M. & Carretta, T. (2021). Development of a situational judgment test to supplement current US air force measures of officership. Military Psychology. 1-16. 10.1080/08995605.2021.1997500.

BPS (2015), The Design and Delivery of Assessment Centres: A Standard produced by the British Psychological Society's Division of Occupational Psychology. Retrieved from: https://www.bps.org.uk/division-occupational-psychology/assessment-standards

Bramble, W. J., & Koonce, J. M. (1998). The Path to Airline Employment: Flight Experience and Performance in a Full-Mission Flight Simulation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 42(11), 797–800. https://doi.org/10.1177/154193129804201106

Breaugh, J.A. (2009). The use of biodata for employee selection: Past research and future directions. Human Resource Management Review, 19, 219-231.

Brehmer, B. (1995). Psychometric analysis of a situational interview for the selection of air traffic controller candidates, Norrköping: Luftfartsverket, MRU delrapport No 18

Brehmer, B. (1998). The MRU situational interview Mark II. An analysis. Norrköping: Luftfartsverket, MRU delrapport No 29

Broach, D. (2012). Incremental validity of biographical data in the prediction of en route air traffic control specialist technical skills (Report No. DOT/FAA/AM-12/8). Washington, DC: FAA Office of Aerospace Medicine.

Broach, D., Schroeder, D., and Gildea, K. (2019). Best Practices in Pilot Selection. Report DOT/FAA/AM-19/06. Washington, DC: FAA Office of Aerospace Medicine.

Campion, M. A., Palmer, & Campion. (1997). A review of structure in the selection interview. Personnel Psychology, 50, 655-702.

Carretta, T. R. (1992). Short-term retest reliability of an experimental U.S. Air Force pilot candidate selection test battery. International Journal of Aviation Psychology, 2(3), 161-173.

Carretta, T. R., & Ree, M. J. (1995). Air force officer qualifying test validity for predicting pilot training performance. Journal of Business and Psychology, 9(4), 379-388.

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

Carretta, T.R. and Ree, M.J. (1997a) Expanding the nexus of cognitive and psychomotor abilities. International Journal of Selection and Assessment, 5, 149-158.

Chaiken, S.R., Kyllonen, P.C. & Tirre, W.C. (2000). Organization and components of psychomotor ability. Cognitive Psychology, 40, 198-226.

Christian, M., Edwards, B., & Bradley-Geist, J. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. Personnel Psychology. 63. 83 - 117. 10.1111/j.1744-6570.2009.01163.x.

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). Psychological testing and assessment: An introduction to tests and measurement (3rd ed.). Mayfield Publishing Co.

Collins, W. E., Nye, L. G., and Manning, C. A. (1990). Studies of poststrike air traffic control specialist trainees: III. Changes in demographic characteristics of academy entrants and biodemographic predictors of success in air traffic controller selection and academy screening (Report No. FAA-AM-90-4). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Connelly, E. M., & Shipley, B. D. (1982). Development of sensitive performance measures for selection of crews for flight training. Paper presented at the Annual Meeting of the Human Factors Society.

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, p. 240–252.

Corstjens, J., Lievens, F. and Krumm, S. (2017). Situational Judgement Tests for Selection. In The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention (eds H.W. Goldstein, E.D. Pulakos, J. Passmore and C. Semedo).

Cristofori, I., Cohen-Zimerman, S. & Grafman, J. (2019). Executive functions. In Handbook of Clinical Neurology 163:197-219. https://doi.org/10.1016/B978-0-12-804281-6.00011-2.

Dailey, J.T., & Pickrel, E.W. (1984). Development of the Air Traffic Controller Occupational Test. In Sells, Dailey and Pickrel (Eds.) Selection of Air Traffic Controllers. Washington, DC: Federal Aviation Administration.

Damitz, M., Chetcuti, F. & Henriques, P. (2010). Validation of the First European Air Traffic Controller Selection Tests (FEAST). Technical report, Eurocontrol, Brussels.

Damitz, M., Manzey, D., Kleinmann, M., & Severin, K. (2003). Assessment center for pilot selection: Construct and criterion validity and the impact of assessor type. Applied Psychology: An International Review, 52(2), 193–212. https://doi.org/10.1111/1464-0597.00131

Damos, D.L. (2020). Individual differences in multiple-task performance: Status and issues. Technical report AFCAPS-IR-2020-0004. Air Force Personnel Center Strategic Research and Assessment Branch, Randolph AF Base. https://doi.org/10.13140/RG.2.2.31314.17600.

Damos, D. L. (2011). Knowledge, Skills, Abilities, and Other Characteristics for Military Pilot Selection: A review of the literature. Retrieved from Randolph AFB, Texas:

Damos, D.L. (1996). Pilot selection batteries: Shortcomings and perspectives. International Journal of Aviation Psychology, 6, 199-209.

Davis, W., Koonce, J., Herold, D., Fedor, D., & Parsons, C. (1997). Personality variables and simulator performance in the prediction of flight training performance. Paper presented at the Ninth International Symposium of Aviation Psychology, Columbus, OH.

de Winter, J., Dodou, D., & Mulder, M. (2012). Training Effectiveness of Whole Body Flight Simulator Motion: A Comprehensive Meta-Analysis. The International Journal of Aviation Psychology. 22. 164-183. DOI: 10.1080/10508414.2012.663247.

Dean, M., Russell, C. & Farmer, W. (2002) Non-Cognitive Predictors of Air Traffic Controller Performance. In Eiβfeldt, H., Heil, M., & Broach, A. (Eds) , Staffing the ATM system (p. 59-72) Aldershot (Uk), Ashgate

Dipboye, R. L. (1994). Structured selection interviews: Why do they work? Why are they under- utilized? In N. Anderson & P. Herriot (Eds.), International handbook of selection and assessment (pp. 455–473). New York: Wiley.

Dolgin, D. L., & Gibb, G. D. (1988). A Review of Personality in Aviation Selection (NAMRL Monograph 36). Retrieved from Pensacola, FL:

Dolgin, D. L., & Gibb, G. D. (1989). Personality assessment in aviator selection. In R. S. Jensen (Ed.), Aviation psychology. (pp. 288-320): Gower Publishing Co.

Dye, D. A., Reck, M., & McDaniel, M. A. (1993). The validity of job knowledge measures. International Journal of Selection and Assessment, 1(3), 153–157. https://doi.org/10.1111/j.1468-2389.1993.tb00103.x

EATCHIP (1996). Guidelines for Selection Procedures and Tests for Ab Initio Trainee Controllers. Report No. HUM.ET1.ST04.1000-GUI-01. Brussels, Belgium: EUROCONTROL

EATMP (2002). Selection Tests, Interviews and Assessment Centres for Ab Initio Trainee Controllers: Guidelines for Implementation (Revised). Report HRS/MSP-002-GUI-03-01. Brussels: EUROCONTROL

Eißfeldt, H. (2002). ATC-Specific Tests. In Eiβfeldt, H., Heil, M., & Broach, A. (Eds) , Staffing the ATM system (p. 85-96) Aldershot (Uk), Ashgate

Eißfeldt, H. (2004). Cost saving: the use of biodata to improve selection efficiency in aviation. In Goeters, K.M. (2004). Aviation Psychology: Practice and Research (pp.209-217). Aldershot (Uk), Ashgate.

Fleishman, E. A. (1953). Testing for psychomotor abilities by means of apparatus tests. Psychological Bulletin, 50, 241-263.

Fleishman, E. A. (1964). The structure and measurement of physical fitness. Englewood Cliffs, NJ: Prentice–Hall.

Fleishman, E. A., & Hempel, W. E., (1954). Changes in factor structure of a complex psychomotor test as a function of practice. Psychometrika, 19, 239–252.

Fleishman, E.A. (1953). Testing for psychomotor abilities by means of apparatus tests. Psychological Bulletin, 50(4), 241-262.

Fowler, B. (1981). The aircraft landing test: An information processing approach to pilot selection. Human Factors, 23(2), 129-137.

Fox, K.D. (2011) Prediction of Air Traffic Controller Trainee Selection and Training Success Using Cognitive Ability and Biodata. Walden Dissertations and Doctoral Studies Collection

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72(3), 493–511. https://doi.org/10.1037/0021-9010.72.3.493

Goerke, P. & Maier, J. (2022). Analysis of a situational judgment test for teamwork as a preselection tool for an assessment center: A construct-based approach. International Journal of Selection and Assessment.1-9. DOI: 10.1111/ijsa.12391

Goldstein, G., Beers, S. R. & Herse, M. (2004). Comprehensive Handbook of Psychological Assessment. Wiley Online Library.

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

Hays. R.T., Jacobs, J.W., Prince, C., & Salas, E. (1992). Flight Simulator Training Effectiveness: A Meta-Analysis, Military Psychology, 4:2, 63-74, DOI: 10.1207/s15327876mp0402_1

Hättig, H.J. (1991). Selection of Air Traffic Control Cadets. In R. Galand, and A.D. Mangelsdorff (Eds.), Handbook of Military Psychology (p. 115-129). New York, NY: John Wiley & Sons.

Hermelin, E., Lievens, F., & Robertson, I. (2007). The Validity of Assessment Centres for the Prediction of Supervisory Performance Ratings. International Journal of Selection and Assessment 15. 405-411. 10.1111/j.1468-2389.2007.00399.x.

Hilton, T. F., & Dolgin, D. L. (1991). Pilot selection in the military of the free world. In R. Gal & A. D. Mangelsdorff (Eds.), (pp. 81-101). Chichester: John Wiley and Sons.

Hörmann, H., Radke, B., & Höft, S. (2007). Measurement of social competence in pilot selection. 14th International Symposium on Aviation Psychology, Dayton, OH p. 263-267.

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

Hörmann, H.-J., Radke, B., & Hoeft, S. (2007). Measurement of social competence in pilot selection. Paper presented at the International Symposium on Aviation Psychology, Dayton, OH.

Hoffman, B.J., Kennedy, .C.L, LoPilato, A.C., Monahan, E.L., & Lance, C.E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. Journal of Applied Psychology, Jul;100(4):1143-68. doi: 10.1037/a0038707.

Höft, S. & Pecena, Y. (2004). Behaviour-oriented evaluation of aviation personnel: an assessment center approach. In Goeters, K.-M. (Ed.), Aviation Psychology: Practice and Research (pp. 153-170). Hampshire, England: Ashgate Publishing Ltd.

Huffcutt, A. I., Conway, J. M., Roth, P. L. & Stone, N. (2001). Identification and Meta-Analytic Assessment of Psychological Constructs Measured in Employment Interviews. The Journal of Applied Psychology. 86. 897-913. 10.1037//0021-9010.86.5.897.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U. C. (2004). Evaluation and comparison of the situational and behavior description interview formats. International Journal of Selection and Assessment, 12, 262–273.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology: Employment interview validity. International Journal of Selection and Assessment, 22(3), 297–309. https://doi.org/10.1111/ijsa.12078

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

Hunter, D.R. & Burke, E.F. (1995). Pilot selection. Aldershot: Avebury Aviation.

International Air Transport Association (IATA). (2019). Guidance material and best practices for pilot aptitude testing (3rd Ed.). Retrieved from https://www.iata.org/en/publications/pilot-testing

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. International Journal of Selection and Assessment, 17(3), 243–253.

International Test Commission (2001). International Guidelines for Test Use, International Journal of Testing, 1(2), 93-114. https://www.intestcom.org/files/guideline_test_use.pdf

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. Journal of Applied Psychology, 67, 577–580.

Jensen, A.R. (1982) Reaction time and psychometric g. In H.J. Eysenck (ed.) A Model for Intelligence. New York: Springer

Koonce, J. M. (1979). Predictive Validity of Flight Simulators as a Function of Simulator Motion. Human Factors, 21(2), 215–223. https://doi.org/10.1177/001872087902100209

Latham, G. P., & Sue-Chan, C. (1999). A meta-analysis of the situational interview: An enumerative re- view of reasons for its validity. Canadian Psychology, 40, 56–67.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. Journal of Applied Psychology, 65, 422–427.

Lievens, F., Peeters, H., and Schollaert, E. (2008). Situational judgment tests: A review of recent research. Personnel Review. 37, (4), 426-441. Research Collection Lee Kong Chian School Of Business

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. The Journal of Applied Psychology, 96 5, 927-40

Martinussen, M. (1996). Psychological Measures As Predictors of Pilot Performance: A Meta-Analysis. International Journal of Aviation Psychology, 6(1), 1-20.

Martinussen, M., & Torjussen, T. (1998). Pilot Selection in the Norwegian Air Force: A Validation and Meta-Analysis of the Test Battery. International Journal of Aviation Psychology, 8(1), 33-45.

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

Maschke, P. (2004). Personality evaluation of applicants in aviation. In K.-M. Goeters (Ed.), Aviation Psychology: Practice and Research (pp. 141-151). Burlington, VT: Ashgate.

McCrae, R. R., & Costa, P. T., Jr. (1989). More reasons to adopt the five-factor model. American Psychologist, 44(2), 451-452.

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. American Psychologist, 57(5), 509-516.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. Journal of Applied Psychology, 86(4), 730–740. doi:10.1037/0021-9010.86.4.730

McDaniel, M. A., & Nguyen, N. (2003). Situational Judgment Tests: A Review of Practice and Constructs Assessed. International Journal of Selection and Assessment. 9. 103 - 113. 10.1111/1468-2389.00167.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. Personnel Psychology, 60(1), 63–91. DOI: 10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis. Journal of Applied Psychology, 79(4), 599-616

Meierfrankenfeld, K., Gress, W., & Vorbach, t. (2015). Simulator-based assessment of flight-specific aptitudes in German Armed Forces aircrew selection Paper presented at the 18th International Symposium on Aviation Psychology, Dayton OH. pp. 620-623.

Melton, A. W. (1947). Apparatus tests. Army Air Forces Aviation Psychology Program Research Reports, No. 4. Washington, D.C.: U.S. Government Printing Office.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than g. Journal of Applied Psychology, 79(6), 845–851. https://doi.org/10.1037/0021-9010.79.6.845

Paullin, C., Katz, L., Bruskiewicz, K., Huston, J., & Damos, D. (2006). Review of Aviator Selection (TR 1183). Retrieved from Arlington, VA:

Pierce, L. G., Broach, D., Byrne, C. L., and Bleckley, M. K. (2014). Using Biodata to Select Air Traffic Controllers. (Report No. DOT/FAA/AM-14/8). Washington, DC: FAA Office of Aerospace Medicine.

Rathje, H. (2002). Test Criteria – Principles and Standards. In Eiβfeldt, H., Heil, M., & Broach, A. (Eds), Staffing the ATM system (p. 27-44) Aldershot (Uk), Ashgate

Reber, A., Reber, E., & Allen, R. (2009). The Penguin Dictionary of Psychology (4th edition). London, UK: Penguin Group.

Roth, P.L., Bobko, P., & McFarland, L.A. (2005). A Meta-Analysis of Work Sample Test Validity: Updating and Integrating Some Classic Literature. Personnel Psychology, Winter 2005, 58(4), 1009-1035.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? Journal of Applied Psychology, 75(2), 175-184. DOI: 10.1037/0021-9010.75.2.175.

Rippon, T. S., & Manuel, E. G. (1918). Report on the essential characteristics of successful and unsuccessful aviators. Lancet (September 28), 411-415.

Salgado, J.F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. European Journal of Work and Organizational Psychology, 11 (3), 299-324.

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

Signori, E. I. (1949). The Arnprior experiment: A study of World War II pilot selection procedures in the RCAF and RAF. Canadian Journal of Psychology, 3, 136-150.

Society for Industrial and Organizational Psychology (SIOP). (2018). Principles for the Validation and Use of Personnel Selection Procedures. (5th ed.). Bowling Green, OH: Author.

Spinner, B. (1989). Using the Canadian Automated Pilot Selection System to predict performance in primary flying training: Straight and level flight. Paper presented at the Fifth International Symposium on Aviation Psychology, Columbus, OH.

Spinner, B. (1991). Predicting Success in Primary Flying School from the Canadian Automated Pilot Selection System: Derivation and Cross-Validation. International Journal of Aviation Psychology, 1(2), 163.

Stokes, G. S., Mumford, M. D., & Owens, W. A. (Eds.). (1994). Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction. CPP Books.

Stricker, L. J. (2005). The biographical inventory in naval aviation selection: Inside the black box. Military Psychology, 14(1), 55-67.

Sullivan, T. S., Burgoyne, T. C., McCloy, R. A., & Whetzel, D. L. (2019). Review of situational judgment test (SJT) prototype development process (AFRL-RH-WP-TR-2019-0058). Wright-Patterson AFB, OH: 711 Human Performance Wing, Warfighter Interface Division, Collaborative Interfaces and Teaming Branch. U.S. Air Force Academy. (2020). Military performance appraisal (USAFI 36-2401). https://apps.dtic.mil/sti/pdfs/AD1083950.pdf

Swiss Assessment (2004). Quality Standards set by Swiss Assessment for the development implementation, and evaluation of assessment centres. Retrieved from: https://www.swissassessment.ch/wp-content/uploads/ENG_Standards-lange-Version_21-02.pdf

Tett, R. P., Jackson, D. N. & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta- analysis. Personnel Psychology, 44, 703-720.

The British Psychological Society (2017). Psychological testing: A test user's guide. Leicester, UK: The British Psychological Society.

Thornton, G. C. III, & Rupp, D. E. (2006). Assessment centers in human resource management: Strategies for prediction, diagnosis, and development. Lawrence Erlbaum Associates Publishers.

Trites, D.K. (1961). Problems in Air Traffic Management I. Longitudinal Prediction of Effectiveness of Air Traffic Controllers. Report No. DOT/FAA/AM-61-1. Oklahoma City, OK: FAA Civil Aeromedical Research Institute

Turnball, G. J. (1992). A review of military pilot selection. Aviation, Space and Environmental Medicine, 63(9), 825-830.

Viteles, M. S. (1945). The Aircraft Pilot: 5 Years of Research A Summary of Outcomes. Psychological Bulletin, 42(8), 489-526.

Walters, L. C., Miller, M. R., & Ree, M. J. (1993). Structured Interviews for Pilot Selection: No Incremental Validity. International Journal of Aviation Psychology, 3(1), 25-38.

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

Whetzler, D.L., & McDaniel, M.A. (2009). Situational judgment tests: An overview of current research. Human Resource Management Review, 19(2), 188-202.

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

Zierke, O. (2011). Knowledge Tests for Ab-Initio Pilot Selection in Civil Aviation. 16th International Symposium on Aviation Psychology, 287-291. https://corescholar.libraries.wright.edu/isap_2011/67

Zierke, O. (2014). Predictive Validity of Knowledge Tests for Pilot Training Outcome. Aviation Psychology and Applied Human Factors 4(2), 98-105.

Zinn, F., Goerke,P. & Marggraf-Micheel, C. (2020). Selecting for cockpit crew. In Bor, R., Eriksen, C., Hubbard & King, R. (Eds), Pilot Selection -Psychological Principles and Practise (p.21-34), CRC Press, New York.

7. Design and Utility of Selection Systems

Jennifer Eaglestone & Karien Stadler

A selection procedure refers to a composite of the different stages of selection used to select candidates for a certain position. It can start with pre-screening and include several rounds of assessment, and, for some candidates, it ends with a job offer.

The development of a robust, effective and efficient selection procedure has been well documented in the literature, see for instance Roe (1998, 2005, 2017). For aviation specific information please see for example Guidance Material and Best Practices for Pilot Aptitude Testing (IATA, 2019), Guidelines for Selection Procedures and Tests for Ab Initio Trainee Controllers (EUROCONTROL, 2001) and Selection tests, interviews and assessment centres for ab initio trainee controllers: Guidelines for implementation (Eurocontrol, 2002).

This chapter provides a short guide to designing and implementing your selection procedure, touching on the most important considerations. For more detailed information one must refer to the various references above, as well as those contained in the rest of the chapter.

7.1. Designing Your Selection System

This section provides an introduction to designing a selection system. The content covered in chapters 5, 6 and 9 are useful resources, when considering the identification of KSAOs and instruments, and the validation of selection procedure as a whole. Here we discuss the job requirement profile, instrument selection and the choice of selection procedure structure.

7.1.1. Defining Your Job Requirements Profile

The first step in designing a selection system is to define the job requirement profile for the position/role you are selecting for by using the results from a job analysis. As discussed in chapter 5, a common framework for creating a job requirement profile is to use KSAO's (Knowledge, Skills, Abilities, Other). Subsequently, with your job requirement profile in hand, these requirements need to be linked to training success and/or effective job performance.

In some cases, the results of a job analysis may be too expansive, or too ambiguous, to be clearly linked to job requirements. It is in these cases that the practitioner (for instance by following a more focused follow-up analysis and or consulting with SMEs) must decide which attributes to focus on when selecting new individuals for the role. Ideally selection should focus on those characteristics that cannot be trained (within the time available) and that varies widely within the applicant pool (Hörmann, 1998a). The number of KSAO's should also be kept to a level that can be effectively assessed during the process (BPS, 2015).

7.1.2. Choosing Selection Methods & Instruments

Once you have decided which KSAOs you will be selecting for, the next step in the process is to choose your selection methods and instruments (for more information and instrument criteria see chapter 6). As mentioned in chapter 6, psychometric criteria (objectivity, reliability, validity and utility) are important to take into consideration when choosing your instruments. The validity of certain types of instruments for predicting performance in training and on the job, varies widely and when selecting instruments one must also examine the degree to which a second predictor has incremental validity over the first (see chapter 6).

Social and practical criteria such as applicability, availability, face validity and fairness should also be taken into consideration when designing your selection procedure. These criteria can be considered complimentary to the predictive validity and besides the effect they have on your selection system itself, they can also affect the attractiveness of your organisation to potential applicants (Hörmann, 1998c).

Lastly it is important to mention that when choosing your instruments, it is generally considered best practice to measure each KSAO at least twice within your selection procedure (but not so many times that measurements become redundant). Every instrument has its own bias of measurement, thus to reduce the possibility that your outcome is based on errors of measurement at least two independent measures should be included (Hörmann, 1998c).

Your choice of selection instrument per KSAO can be recorded in a so-called selection matrix. An example of a selection-matrix can be found in Table 7.1.

KSAO	Test 1	Test 2	Test 3	Job Sample	Interview	Etc.
Perceptual Speed		x		x		
Vigilance	x			x		
Decisiveness			x	x	x	
Etc.						

Table 7. 1: Example of part of a selection matrix

7.1.3. Choosing a Structure

There are two common structures used for selection procedures. The most basic of the two is a single-stage process in which all candidates are assessed using all selection methods. This method is usually only put in place when getting a candidate to the testing location is difficult or costly, when only a few selection instruments are administered (Weissmuller & Damos, 2014) or when there are very few candidates.

The second common structure is the multiple hurdle (also known as the multistage or funnel) approach (see figure 7.1 for a schematic representation). In this structure only those who pass a selection round can take part in the subsequent stages of selection (with the others being rejected). If a job profile includes a large range of KSAOs and these need to be assessed for a large group of applicants, then this structure is generally more efficient (Hörmann, 1998a).

A multi-stage approach can result in significant savings in selection costs, as unpromising candidates are rejected early in the process and the most expensive selection methods such as simulator assessments and interviews can be administered in the final stages of the process for fewer candidates (Weissmuller & Damos, 2014). The cost issues related to a selection procedure will be discussed in more detail in section 7.3. For some examples of multiple hurdle approach selection systems in aviation see Goeters (1998).

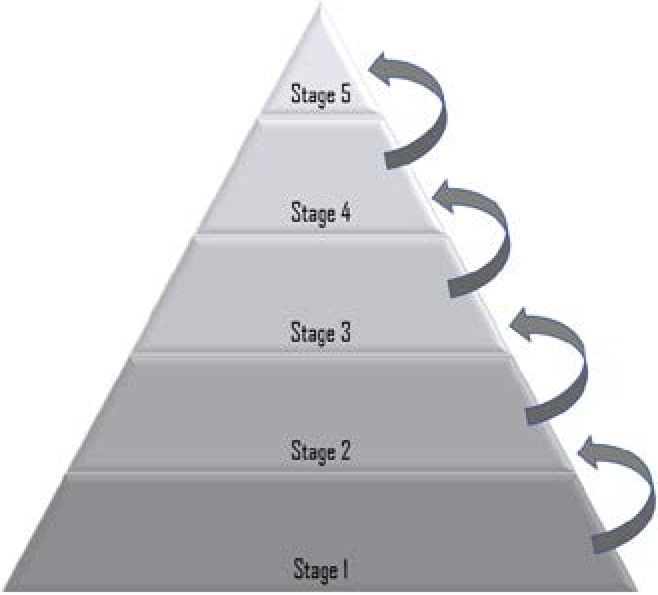

Figure 7. 1: A multiple-hurdle approach (schematic)

7.1.4. Decision Strategies

After having candidates take part in the various stages of selection the goal is to find those candidates that are most suitable for the job. Whichever structure you have chosen there will be several moments at which decisions need to be made – on whether a candidate should go through to further testing or if someone should be offered a job or position in training. This section discusses the various decisions. We discuss two levels of decision-making, the first involves decisions per selection round involving cut-off scores and/or ranking. The second is about the strategy on how to combine all the information collected about an applicant and make the final job offer or not.

7.1.5. Competitive Ranking & Cut-off Scores

If you have chosen a multiple hurdle approach for your selection, then after each stage of selection decisions must be made on who will go through to the next round. This can be done using cut-off scores and, or, a method of ranking.

When setting cut-off scores, various methods can be considered. Cut-off scores can be set according to an empirically defined cut-off point or score distribution. It can be set to reduce false positive or false negative errors (see 7.3 for more information) or it can be used as advised by the specific test provider. The more stringent you are, the more challenging it becomes to find a reasonable number of applicants meeting the requirements. In a comprehensive multi-stage process, one option could be to be more lenient in terms of setting cut-off scores in the initial screening phase and more rigorous when setting cut-off scores later in the process. Another option could be to be more stringent when setting cut-off scores during the initial stages to reduce selection cost but could result in a situation where you do not select enough candidates to meet the selection target (see chapter 8.1.1 for further information on range restriction issues).

When implementing cut-off scores for cognitive ability tests, there is the potential for adverse impact as group differences may be observed because of factors other than competence (for example differences in educational systems). For more detail on how to minimize subgroup differences when designing a selection process, see Outtz (2009).

Ranking or top-down selection refers to the practice of offering available positions in rank order from highest to lowest, extending offers until all positions are filled. (Gatewood & Field, 2001). If ranking it is important to still apply a minimum score (cut-off) to ensure candidates have at least the minimum competence required.

7.1.6. Combining Selection Data

In any sophisticated selection process, a great deal of information is collected and recorded in the different stages of selection. There are two common approaches to combine selection data, mechanical (statistical) decision-making and the clinical (expert) approach. Both the single-stage and the multiple hurdle process can be implemented with or without a statistical model (Weissmuller & Damos, 2014).

The mechanical (statistical) decision-making approach involves using statistical techniques to make predictions where different test results are for instance mathematically combined to predict success in training or on the job. These can include multiple regression & multiple cut-offs (see Guion, 1998, for more information on mechanical decision-making using these methods) and more recently, predictive modelling techniques using machine learning (see for example Drasgow & Olson-Buchanan, 2021). See Meijer et. al. (2020) for further explanation on how to use mechanical decision-making in selection.

The second and more popular approach is the clinical method (also known as the holistic, or expert judgement method) which can include both individual judgements of data as well as group consensus meetings. Here decision-makers look at test results from each round of selection and then make a judgement on the success of the candidate (Born & Scholarios, 2005). Decisions must be taken by an interdisciplinary team of highly qualified professionals taking the strengths and weaknesses of the instruments into consideration as well as the present and future demands of the job (Hörmann, 1998a). For aviation specific guidelines on clinical decision-making in selection see for instance Hörmann (1998b).

Research, however, does highlight the superiority of mechanical decision-making in selection. Grove et al. (2000) carried out the first meta-analysis focusing on the comparison between mechanical and clinical decision-making. Although not purely selection orientated, overall mechanical outperformed clinical decision-making. Kuncel et al., 2013 conducted a meta-analysis to examine the relative predictive power of combining selection data by mechanical methods or holistic methods. Their results indicate that mechanical approaches substantially outperform clinical combination methods and the accuracy of the prediction of job performance was improved by more than 50% when data was combined mechanically rather than holistically. Both methods can of course have bias, but the nature of mechanical decision-making makes detecting and correcting bias easier.

The strong preference for expert judgment we see in practice, could come from the shortcomings of statistical models (Hörmann, 1998b), especially in the past where aviation psychologists were limited to statistical techniques such as multiple regression. Limitations here for instance included that only one job-criterion could be predicted at a time and that multiple regression assumes a linear relationship between test scores and the criterion (Hörmann, 1998b). Even though clinical decisions seem inferior to mechanical decisions it is possible to somewhat enhance the predictive power of clinical selection decisions using mechanically combined data as an anchor and make limited adjustments. Documenting deviations from mechanically combined scores makes the decision public and permits follow up research and feedback. A second possibility is to present both expert and mechanically combined scores to decision-makers (see for instance Roe, 1998, for an example of a semi clinical selection method). Predictive power could also be improved by averaging across multiple raters even if secondary (and possibly less involved) raters were given a lower weight in the final assessment (Kuncel., et al. 2013).

However, you make your selection decisions, they must be based on pre-defined clear-cut and exhaustive set of decision rules as these rules are an essential part of your selection procedure. In fact, one could say that the selection instruments are purely technical components of a greater decision-making system (EUROCON-

TROL, 2002) and therefore not only selection instruments should be validated but also the outcomes of decisions. For more information on validation see chapter 8.

7.2. Utility

Selection is useful when it helps to make decisions and usefulness can therefore be evaluated in terms of correct decisions made. This utility (the practical value of a selection procedure) does not only depend on the validity of the selection procedure but also on the selection ratio and the base rate (Taylor & Russel, 1939).

The selection ratio is the number of available job openings divided by the number of applicants and the base rate is the proportion of people who were judged successfully during the selection procedure. The base rate and the selection ratio both affect the success ratio (the people who were successfully judged divided by the number of applicants), which can in turn be used as a measure of the utility of the selection procedure. (Cascio, 1998). With more applicants successfully judged the practical value of the selection procedure rises and with a rise in the proportion of correct decisions there are usually also quite large economical gains (Schmidt et al., 2016).

By the nature of a selection procedure there will, aside correct acceptances (true positives) and correct rejections (true negatives) also be incorrect acceptances (false positives) and incorrect rejections (false negatives). This concept was first introduced by Taylor & Russel (1939). False positives are often considered worse in selection than false negatives and if they are a major concern then the cut-off score may be raised. If false negatives are more problematic (for instance due to shortage of candidates) then the cut-off score might be lowered (Cascio, 1998).

In figure 7.2 this idea is presented schematically based on Tayor & Russell (1939, p67). The y-axis represents a test-score or composite score and the X-axis represents job performance. What becomes obvious with this chart is that correct and incorrect acceptances and rejections are always related. Lowering the cut-off to reduce the number of incorrect rejections (false negatives) for example, automatically increases the number of incorrect acceptances (false positives).

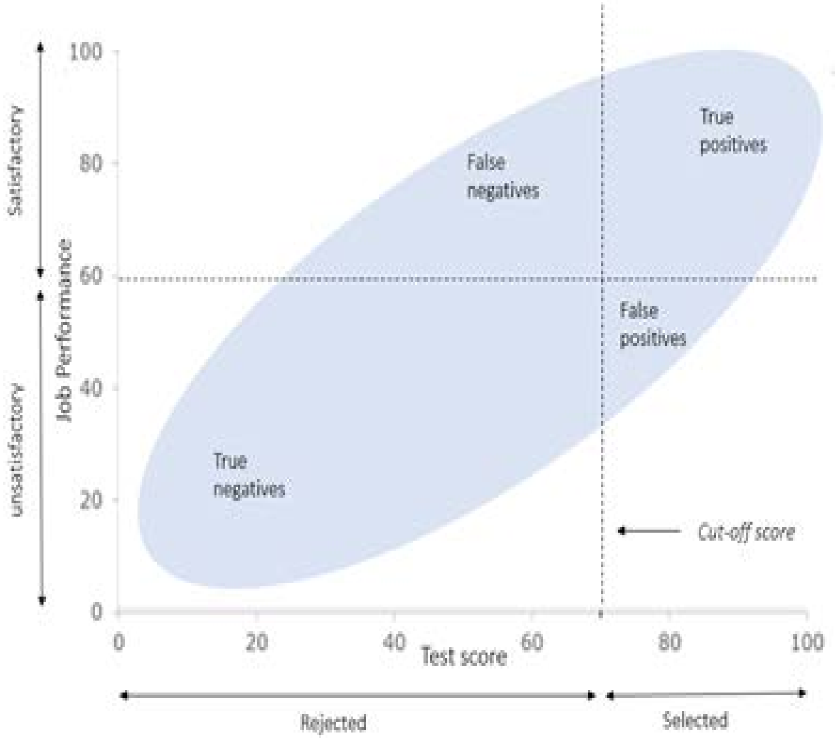

Figure 7. 2: A Taylor-Russell Chart

The variability of job performance also has an effect on utility (Schmidt et al. 2016). If variability in job performance is low then all applicants would perform at the same level if hired. If it is high then it's important to hire those who perform best, and the practical utility of a valid selection procedure is then also very high.

According to Goeters (1998) utility in selection of aviation personnel is mainly found in aspects such as the financial advantage of a low failure rate during ab initio training, benefits in saving costs in recurrent training and safety improvements. He also states that utility for the applicant is found in benefits such as the risk of false personal investment, increasing job-security and promotion of job satisfaction and mental health.

For more information on the utility of selection methods please use all the references above for your further reading.

7.3. Summary

This chapter provides a brief overview of some of the most important factors to consider when designing a selection process. It aims to highlight best practice principles with specific reference to a science-based process. It further provides guidance on setting up your selection procedure and addresses matters to consider when making selection decisions. For more information on problems and pitfalls in selection please see chapter 10. Once implemented a selection process should be evaluated and validated as soon as possible, for more information on validation see chapter 8.

It is important to be aware that these topics are much broader than presented here and when designing or evaluating a selection procedure it is advisable to delve into the general selection literature (for instance Evers et al. (2005), Drenth et al.(1998)) to gain more insight into these topics and the science behind them.

7.4. References

Born, M. P., & Scholarios, D. (2005). Decision making in selection. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), The Blackwell handbook of personnel selection. Malden, MA: Blackwell Publishing

BPS (2015), The Design and Delivery of Assessment Centres: A Standard produced by the British Psychological Society's Division of Occupational Psychology. Retrieved from: https://www.bps.org.uk/division-occupational-psychology/assessment-standards

Cascio, W.F. (1998). Applied Psychology in Human Resource Management 5th ed. USA: Prentice Hall

Drenth, P. J. D., Thierry, H., & de Wolff, C. J. (Eds.) (1998). Handbook of Work and Organizational Psychology, 2nd ed., Vol. I - IV. Psychology Press.

Drasgow, F., & Olson-Buchanan, J.B. (2021). Technology-Driven Developments in Psychometrics. In Scott, J.C., Bartram, D., and Reynolds, D.H. (Eds), Next Generation Technology-Enhanced Assessment: Global Perspectives on Occupational and Workplace Testing (Educational and Psychological Testing in a Global Context). Cambridge: Cambridge University Press. doi:10.1017/9781316407547

EFPA Board of Assessment Document 110c. EFPA review model for the description and evaluation of psychological and educational tests: Test review form and notes for reviewers. Version 4.2.6. 2013

EUROCONTROL (2001). Guidelines for selection procedures and tests for ab initio trainee controllers (Revised). European Air Traffic Management Programme. Document identifier HRS/MSP-002-GUI-01. EUROCONTROL Headquarters, Brussels.

EUROCONTROL (2002). Selection tests, interviews and assessment centres for ab initio trainee controllers: Guidelines for implementation (Revised). European Air Traffic Management Programme. Document identifier HRS/MSP-002-GUI-03-01. EUROCONTROL Headquarters, Brussels.

Evers, A., Anderson, N. & Voskuijl, O. (2005.), The Blackwell Handbook of Personnel Selection. Malden, MA: Blackwell Publishing.

Gatewood, R. and Field, H. (2001) Human Resource Selection. Harcourt Brace & Company, Orlando.

Goeters, K.-M. (1998). Aviation psychology: A science and a profession. Aldershot: Ashgate.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment, 12(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Guion R.M. (1998). Assessment, measurement, and prediction for personnel decisions. Mahwah, HJ: Lawrence Erlbaum.

Hörmann, H-J. (1998a) Basic Concepts. In Goeters, K.-M. (Ed). Aviation psychology: A science and a profession. Aldershot: Ashgate.

Hörmann, H-J. (1998b) Diagnostic Decisions. In Goeters, K.-M. (Ed). Aviation psychology: A science and a profession. Aldershot: Ashgate.

Hörmann, H-J. (1998c) Recommendations and Future Perspectives. In Goeters, K.-M. (Ed). Aviation psychology: A science and a profession. Aldershot: Ashgate.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. Journal of Applied Psychology, 98(6), 1060–1072. https://doi.org/10.1037/a0034156

Lawrence, A.D., & Kinney, T.B., (2017). Mobile Devices and Selection. SIOP White Paper, 2017.

Meijer, R. R., Neumann, M., Hemker, B. T., & Niessen, A. S. M. (2020). A tutorial on mechanical decision-making for personnel and educational selection. Frontiers in Psychology, 10, Article 3002. https://doi.org/10.3389/fpsyg.2019.03002

Outtz, J.L., (2009). Adverse Impact: Implications for Organizational Staffing and High Stakes Selection. 1st Edition. SIOP.

Roe, R. A. (1998). Principles, Models and Techniques. Handbook of Work and Organizational Psychology: Personnel psychology, 3, 2.

Roe, R. A. (2017). The Design of Selection Systems: Context, Principles, Issues. The Blackwell Handbook of Personnel Selection, 73–97. doi:10.1002/9781405164221.ch4

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

Schmidt, F. L., Oh, I. S. & Shaffer, J. A. (2016). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 100 Years of Research Findings.

Taylor, H. C. & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. Journal of Applied Psychology, 23(5), 565.

The American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education. The Standards for Educational and Psychological Testing. 2013

Weissmuller, J.J. & Damos, D.L. Improving the Pilot Selection System: Statistical Approaches and Selection Processes. The International Journal of Aviation Psychology, 2014, 24:2, 99-118.

8. Validation

Hans-Jürgen Hörmann & Jóhann Wium

For the essential proof of a selection test's validity, individual differences in test scores (predictors) have to be related to corresponding differences in job performance (criteria). Usually, selection tests are administered to job applicants before they enter job training or the job itself. Initial data on job performance becomes available after a certain time period, for example when the training or parts of it are completed or after a fixed period of time.

When this criterion information is correlated to the predictor test scores, the predictive validity of the selection tests or the entire test battery is assessed. Alternatively, test scores and job performance can be collected at the same point of time, if a sample of job holders volunteer to take a number of selection tests, which then are related to their job performance. This process is called concurrent validation. Both predictive and concurrent validation fall into the overall category "criterion-related validity", which tests the assumption that better test scores correspond to better job performance.

8.1. Validity Issues

In practice, concurrent and (especially) predictive validation studies, are time consuming and involve a number of methodological issues. For example, the need to keep the selection test data identifiable until job performance becomes available can cause issues with the requirements for data protection and data privacy, or the size of the operation can limit sample sizes. A concurrent validity study in easier to conduct because predictor and criterion information is collected at the same time. However, concurrent validity is only an approximation of predictive validity. Two particularly common problems for selection in aviation are range restriction and issues with the job criteria, both covered here below. A broader overview of validity issues in selection of aviation personnel can be found in Carretta and Ree (2000).

8.1.1. Range Restriction

The size of a correlation coefficient depends on the variances of the involved variables. If the variance becomes systematically restricted, the observed correlation shrinks towards zero (e.g., Sacket & Yang, 2000). During selection, candidates are usually filtered out if test scores are below a certain threshold. These candidates are lost for the test validation. In other words, for the applied selection test a systematic restriction of range occurs, which reduces its correlation with the job criteria.

This can apply even to tests not used for selection decisions (e.g., test under development), as its variance can be reduced through the intercorrelation with other predictor tests used in the selection process. In a multi-stage selection process (e.g., only high scorers from one selection step can take the subsequent tests) the problem of variance restriction can accumulate across the stages. With a low selection ratio (e.g., 10% selected candidates), which is quite common in the selection of aviators, the problem of range restriction can become severe and it affects the estimation of predictive as well as concurrent validity.

Since range restriction hides the predictive validity of selection tests and thus the utility of using the respective predictor, statistical correction formulas should be utilized to correct the originally calculated predictor-criterion correlations and make them more comparable across settings. Carretta and Ree (2003) recommend to use the multivariate correction formula provided by Lawley (1943) if more than one predictor test is used. Validation studies with unrestricted samples are rare (for examples see Dubois, 1947; Thorndike, 1947; Hörmann et al., 2018) and all correction formulas would ideally require estimates for predictor variance in an unrestricted sample to be adopted. However, the more complex a selection process is (number of predictors and stages), the more difficult it becomes to apply the corrections correctly.

8.1.2. Criteria of Job Performance

Choosing appropriate job performance criteria against which test scores can be validated is another issue for validation efforts. The scope of possible job criteria in aviation can range from academic scores, instructor ratings, simulator performance, training duration variables to failed checks or even safety events (see Chapter 2). A selection test can have "different validities" depending on the chosen criterion measure. The value of these measures as a validity criterion depends for example on the relevance for the job, the reliability of the gathered information, and on the characteristics of the scoring (e.g., dichotomous or continues scoring, skewness of the distribution). In most cases researchers have to accept partial or preliminary information about job success (e.g., training data) because ultimately relevant performance criteria are not readily available. In order to compensate for errors of measurement of the criteria the correction for attenuation formula should be applied (e.g., Lord & Novick, 1968). How to compensate for distribution bias (e.g., with a pass/fail criterion) is discussed in detail by Lord and Novick (1968, pp 335ff). The effects of poor criterion reliability, using a dichotomous criterion in addition to range restriction add up and contribute to an underestimation of the true predictive validity. Corrections for these artifacts should be used when possible.

Because of these issues, methodologically sound validation studies are hard to find. In the 1980s influential research on meta-analytic methods was published that looked at the mass of accumulated evidence instead of individual studies. Since the work of Hunter and Schmidt (1977, 1998, 2015), some of these methods became known as validity generalization techniques. The advantage of these techniques is that by aggregation of several individual studies the mean validity coefficients may be estimated based on a large number of studies, and the distribution of the validity coefficients can be scrutinized for variation.

If possible, the correlations should be corrected for statistical artifacts (e.g., range restriction) before they are entered in the meta-analysis. The meta-analytic calculations make it possible to estimate the true variation between validity coefficients, including a credibility interval (Schmidt & Hunter, 2015). If the lower 90% credibility value of this interval is above zero, then this validity information could be generalized to other selection contexts as well (Hunter & Burke 1994). Some of the major meta-studies for selection tests of pilots and ATCOs are summarized in the two sections below.

8.2. Meta Studies for Pilot Selection

Over the decades, several attempts have been made to document the predictive validity of pilot selection tests. A summary is provided by Martinussen and Hunter (2018). Altogether four meta-analytic studies could be identified, which are compared in the following section according to aims, scope (inclusion criteria), samples, and main findings. Further studies on specific predictor sets or specific pilot roles can be found in the literature. For example, Campbell, Castaneda and Pulos (2009) reviewed the significance of personality predictors in military pilot selection and Carretta (2013) investigated the predictive validity of the AFOQT for training success of remotely piloted aircraft (see section 3.7).

8.2.1. Damos (1993)

Aim: To compare the validity of multiple-task measures where two or more tasks are performed concurrently with the validity of corresponding single task measures for the prediction of flight performance.

Inclusion criteria: Searches of the Psychological Abstracts database, NTIS, and Proceedings of conferences on Human Factors or Aviation Psychology (incl. (W) EAAP) were conducted. Studies published in English between 1964 and 1992 were retrieved, which included correlations between multiple-task measures and pilot performance.

Studies and sample size: Both civilian and military pilots with and without experience. 14 multiple tasks studies with 6,920 subjects were compared to 12 single task studies with 5,378 subjects. The predictive time interval was between a few days and one year.

Main findings: Both types of measures were significantly predictive for flight performance with multiple-tasks measure demonstrating significantly higher predictive validity than single-task measures. The meta-analytic methods resulted for the multiple-tasks studies in a mean effect size of $d = 0.48$ ($p < .0001$), which corresponds to a (uncorrected) correlation of $r = 0.23$. The mean effect size for the single-task studies was $d = 0.37$ ($p < .0001$), which corresponds to a correlation of $r = 0.18$ (single task). Additionally, two moderators for the validity were identified. With the multiple-task measures prediction was slightly better for civilian as for military pilots and also slightly better for experienced versus student pilots.

8.2.2. Hunter & Burke (1994, 1995)

Aim: To provide a more exhaustive review of the predictive validity of pilot selection tests

Inclusion criteria: Manual and computerized searches between the years 1940 and 1990 in Psychological Abstracts and published bibliographies of the military services, complemented by "promising" sources cited in the studies obtained. Studies predominantly from North America and the UK were retrieved that at least reported a correlation between predictor tests and flight-training criteria. Care was taken not to duplicate studies, which were found in different articles or technical reports.

Studies and sample size: A total of 68 published studies were identified, from which 468 correlations were extracted for a cumulated sample of 437,258 civilian and military (student) pilots.

Main findings: The mean predictive (uncorrected) validities ranged between r = .06 (for education) to r = .34. (for job sample tests). General ability, verbal ability, fine dexterity, age, education and personality did not show generalizable validity for flight performance. Better predictors were quantitative ability, spatial ability, mechanical, aviation information, general information, gross dexterity, perceptual speed, reaction time, biodata inventory, and job sample. As a moderator the decade of publication was significantly associated with a decline in average validities since over the five decades of studies examined. Higher mean validities were found for Air Force studies, for fixed-wing pilots compared to rotary wing pilots, and for non-U.S. studies compared to U.S. studies.

8.2.3. Martinussen (1996)

Aim: The purpose of this study was to review predictive validities of psychological measures used in pilot selection and to examine possible moderators to the validity.

Inclusion criteria: Computer searches in several psychological and medicine databases were conducted with the keywords pilots, test, and validity. This resulted in a broader geographic sample than reported by Hunter and Burke (1994). This was accompanied by personal contacts with international colleagues and the NATO Aircrew Working Group. Criteria for study inclusion were that bivariate correlations were reported together with the number of subjects or alternate figures that could be converted into these. No time-period was specified, but the search included a wider range from the year 1919 to 1993 (median year 1973).

Studies and sample size: 66 independent samples from 50 studies in eleven different nations (26% non-English speaking countries) were identified. The sample sizes varied between 12 and 2,356. The total number of subjects is not reported but can be. estimated between $17,900 < N < 26,000$. Most subjects were military pilots from both fixed wing (67%) and rotary wing aircraft (21%).

Main findings: Corrected for dichotomization the mean predictive validities of the different test categories ranged from $r_{cor} = .14$ (for personality) to $r_{cor} = .30$ (for training experience prior to selection) for a global pilot-performance criterion. The next best predictors after training experience were cognitive and psychomotor tests (both $r_{cor} = .24$). A composite score of several tests reached $r_{cor} = .37$. With respect to other analyzed moderators (pilot performance: pass/fail on theory; aircraft type: fixed or rotary wing; year of publication), a negative correlation between the publication year and the reported validity coefficient can be mentioned.

However, range-restriction effects could have accounted for these differences as well. For all test categories except personality and biographical inventories, the lower end of the credibility interval was above zero, indicating that, even though moderators may be operating, validity generalizes across settings.

8.2.4. AlMamari and Traynor (2019)

Aim: This study meta-analyzed the predictive validity of composite scores from entire test batteries for pilot performance criteria. It updates previous studies by shifting the time period to the recent 30 years beyond 1987.

Inclusion criteria: Electronic searches in databases such as Defense Technical Information Center, Google Scholar, PsychInfo, ProQuest were conducted for research on predictive validities of ability tests in pilot selection. The searches were guided by keywords including "pilot selection", "selection tests", "flight aptitude tests" as well as test batteries such as "AFOQT", "ASTB", "MICROPAT", "CogScreen", and "PILAPT". This was complemented by manual searches in journals on aviation psychology or human factors as well as conference abstracts by the International Military Testing Association. Inclusion criteria were that sufficient information about the dependent and independent variables was provided, that the predictor was a composite score of at least two tests, that correlations could be derived and sample size reported. The years between 1987 and 2018 were covered.

Studies and sample: 118 independent samples with altogether 116,806 cases were extracted from 89 studies. The majority of samples involved military aviation (52 USAF samples16 U.S. Navy, 7 Canadian Forces, 18 other militaries, 19 civilian airlines, 6 samples from university flight programs).

Main findings: The methodology used was the Schmidt and Hunter (2015) psychometric meta-analysis approach. Because of missing information and the nature of the composite scores, the predictor-criterion correlations could not be corrected for range restriction or unreliability. Mean predictive validities for an overall pilot performance index ranged from $r = .10$ to $r = .34$. The flight simulator as a work sample test was the best predictor ($r = .34$), followed by Acquired Knowledge (mostly verbal tests; $r = .19$), and General Ability ($r = .18$). With regard to moderator effects, a negative correlation between the validity coefficients and the year of publication was not confirmed. The only significant moderator was the level of measurement for the criterion. For some test-batteries a continuous criterion could be better predicted than a dichotomous criterion. The flight simulator was the best predictor for practical flying performance.

8.3. Meta Studies for ATCO Selection

Three meta-analysis have been carried out on validation studies for ATCO selection. Unfortunately, one of the studies was unavailable except in tabular form summarizing its results.

8.3.1. Martinussen, Jenssen and Joner (2000)

Aim: This study meta-analyzed 9 distinct types of tests with both training and performance criteria.

Inclusion criteria: Searched for studies from 1952-1999 on PsychInfo, Medline, CAMI database and proceedings from the Symposium on Aviation Psychology (1987-1999) that included data on sample size, bivariate correlations or data that could be converted into correlation coefficients.

Studies and sample size: 25 studies with between 224 and 11,255 subjects.

Main findings: Best predictors of ATC performance were spatial ability (mean $r_{uncor} = .25$) and work samples ($r_{uncor} = .28$) with personality measures as the worst predictor ($r_{uncor} = .03$) although personality measures had been combined into a single global category that could explain the lack of predictive power.

8.3.2. Schemmer et al. (1996)

Aim: An offshoot of the SACHA (Separation and Control Hiring Assessment) project carried out in the US in the late 90s (Nickels et al., 1995). From the data that is available it can be assumed this study was done to identify the requisite ATCO attributes necessary for success. The meta-analysis looks at multiple predictors (basic cognitive ability, experience measures, personality/interest, simulation exercises, multi-tasking, visual/spatial measures, demographic measures, visual reaction time and psychomotor tests) and multiple criteria (attrition, field training process, field training duration, supervisor appraisal, instructor ratings, performance on the FAA Screen, simulator scores and controller skills tests).

Inclusion criteria: Unknown but limited to datasets from the FAA.

Studies and sample size: Looking at every combination of 10 predictor groups and 11 criterion measure groups, it analysed multiple studies (1 – 1,289) each with multiple subjects (ranging from 112 to 515,883).

Main findings: Across all criterion measures the best predictors are: visual reaction time measures (mean sample $r = .24$), spatial/visual ability (mean sample $r = .23$) and psychomotor measures (mean sample $r = .22$).

Notes: The table of results can be found in Table 3.1.3 in Ramos et al. (2001)

8.3.3. Mouratille, Amadieu and Matton (2022)

Aim: This study investigated the connection between cognitive and non-cognitive factors to training success of air traffic controllers.

Inclusion criteria: Authors searched for studies from 1961-2021 on PA PsycINFO, Scopus, Psychology and Behavioral Sciences Collection, Web of Science Core Collection, ProQuest Dissertations and Theses and Google Scholar. Studies included had to use some manner of cognitive, personality, motivation or biodata predictors, provide sample size and effect size information (or provide sufficient data to calculate effect size) and have no correction for restriction of range. Studies also needed to have been done on ab initio students, use initial or OJT criterions and report data once the training had been fully completed.

Studies and sample size: 51 studies with a cumulated sample of 65,839 subjects.

Main findings: Mean true score correlation for an overall cognitive predictor was $p = 0.37$ and $p = 0.15$ for non-cognitive factors. Strongest cognitive predictors (all $p > 0.3$) were quantitative knowledge, processing speed, work samples, short-term working memory, cognitive composite and visual-spatial processing. Strongest non-cognitive predictors (all $p > 0.3$) were non-cognitive composite and education.

8.4. Summary

The reported findings from meta-analytic validation studies in both pilot and ATCO selection have consistently shown that a number of cognitive ability tests (including multiple-task and psychomotor tests) do have reasonable validities for the prediction of training success and job performance. Also, the predictive validity generalizes across settings. However, when it comes to personality measures the results are rather disappointing even though personality characteristics were identified as important job requirements for ATCOs and pilots (see chapter 3 and 4). One possible reason could be the test format. Personality is often measured with self-description questionnaires. In a selection context, self-descriptions might be subject to impression management strategies and lack of self-insight of the applicants, which could degrade construct validity. Helmreich et al. (1986) referred to potential “honeymoon effects” which might cause postponed effects of personality on pilot performance to be detected after initial flying skills and procedures have been overlearned. A different set of criteria and a stronger focus on long-term work performance might lead to better results. Martinussen (1996) and Martinussen and Hunter (2018) underline that job performance and personality constructs can be positively or negatively related and that when averaging correlations across different personality scales this might level out existing relationships.

8.5. References

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

Carretta, T.R. & Ree, M.J. (2003). Pilot selection methods. In P.S. Tsang & M.A. Vidulich (Eds.) Principles and Practice of Aviation Psychology (pp 357-396). Mahwah/NJ: Lawrence Erlbaum Associates.

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

Dubois, P.H. (Ed.) (1947). The Classification Program. Report No. 2. Army Air Forces Aviation Psychology Program Research Reports. Washington, DC: Army Air Forces.

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

Hunter, D.R. & Burke, E.F. (1995). Pilot selection. Aldershot: Avebury Aviation.

Hunter, J.E., & Schmidt, F.L. (1990). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.

Lawley, D.N. (1943). A note on Karl Pearson's selection formulae. Proceedings of the Royal Society of Edinburgh, 62, Section A, Part I, 28-30.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Reading MA: Addison-Wesley Publishing Company.

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

Martinussen, M. & Hunter, D.R. (2018). Aviation psychology and human factors (2nd ed). Boca Raton: CRC Press Taylor & Francis Group.

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

Mouratille, D., Amadieu, F., & Matton, N. (2022). A meta-analysis on air traffic controllers selection: cognitive and non-cognitive predictors. Journal of Vocational Behavior, Volume 138. https://doi.org/10.1016/j.jvb.2022.103769.

Nickels, B. J., Bobko, P., Blair, M. D., Sands, W. A., & Tartak, E. L. (1995). Separation and Control Hiring Assessment (SACHA) final job analysis report. University Research Corporation.

Ramos, R., Heil, M., & Manning, C. A. (Eds.) (2001a). Documentation of the validity of the AT-SAT computerized test battery: Volume I. Report No. DOT/FAA/AM-01/5. Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Schemmer, F.M., Cooper, M.A., Blair, M.D., Barton, M.A., Kieckhaefer, W.F., Porter, D.L., Abrahams, N. Huston, J. Paullin, C., & Bobko, P. (1996). Separation and Control Hiring Assessment (SACHA) Interim Approach Predictors Volume 1: Technical Report. Bethesda, MD: University Research Corporation.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529–540.

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

Schmidt, F. L., & Hunter, J. E. (2015). Methods of meta-analysis: Correcting error and bias in research findings (3rd ed.). Newbury Park, CA: Sage Publications. DOI: https://dx.doi.org/10.4135/9781483398105

Thorndike, R. L. (1947). Research problems and techniques (Report No. 3). Washington DC: AAF Aviation Psychology Program Research Reports, U.S. Government Printing Office.

9. Current Practices

Hans-Jürgen Hörmann

The current practices for selection of aviation personnel have been reviewed and documented in the past by IATA, as the international organization representing worldwide 290 airlines and EUROCONTROL, as the European organization for the safety of civil and military air navigation with currently 41 member states. The applied selection tools should be compliant with the relevant national and international regulatory material (e.g., EASA 2019) and industry standards provided by professional organizations such as the International Test Commission (ITC), The European Federation of Psychologists' Associations (EFPA), or the American Psychology Association (APA). The following recommendations are based on these references.

9.1. Recommendations from International Aviation Organizations

EUROCONTROL (2001b) proposed that national air-traffic service providers should implement a structured process, called "pre-selection" to determine the eligibility of applicants for the actual ATCO-selection. This process aims to identify applicants who have a higher probability of passing subsequent formal selection stages. The general principle of such a procedure is to define a profile of those characteristics (e.g., education, grades, achievements, goals, specific experience, or medical criteria) that discriminate between accepted and rejected candidates during selection. The advantage of such an extra stage is that it allows for a higher selection rate without increasing the percentage of failures in training. However, this is based on the assumption that the applied selection system has proven to have predictive validity for training and job performance criteria. Candidates' information about pre-selection criteria can be collected by tailored application forms and/or biographical questionnaires as described in chapter 6.4.2. Another option is to publish pre-selection criteria directly in job posts or on the organization's career websites in order to allow potential applicants to conduct a self-assessment whether their own profile matches those criteria. How an occupation or occupational training can be marketed to potential applicants is described for the example of ATCOs by EUROCONTROL (2000, 2003).

In pilot selection such filters are usually not as stringent as those recommended for ATCOs by EUROCONTROL. Generally, basic eligibility criteria (e.g., license requirements, age, level of education) are checked before a candidate is invited to the selection procedure. Legal criteria for a psychological assessment of pilots have recently also been defined. In 2018, EASA decided that all flight crew members shall go through a (non-clinical) psychological assessment before commencing line-flying (Commission Regulation 2018/1042). The respective AMC published in 2019 (AMC1 CAT.GEN.MPA.175(b)), that came into effect in February 2021, specified the following criteria to be assessed:

- cognitive abilities;
- personality traits;
- operational and professional competencies; and
- social competences in accordance with crew resource management principles

The operator has to ascertain by a job analysis that the applied assessment criteria relate to the complexity and the challenges of the operational environment that the flight crew is likely to be exposed to. It is furthermore the operator's responsibility to establish a policy which assures that the assessment is in compliance with (inter-)national codes of conduct for psychological testing as well as with recognized industry standards and best practices in the field of pilot selection. International standards for psychological testing and assessment have been published for example by the International Test Commission (ITC, 2013), the European Federation of Psychologists' Associations (EFPA, 2013) or the American Psychological Association (APA, 2017, 2018).

As a general guidance for pilot aptitude testing, EASA recommends the IATA manual (IATA, 2019), which suggests a number of psychological dimensions to be considered in pilot selection. The dimensions of the IATA Pilot Aptitude Testing (PAT) Matrix are seen as essential predictors for the ICAO Competency Framework for pilots, as proposed in the approved amendments to Annex 1 and the PANS-TRG, Doc 9868, applicable since 5th November 2020. IATA also suggests assigning differential weights to the dimensions in correspondence to the level of experience of the applicant from ab-initio student pilots to licensed pilots in command as described in chapter 7.

Job analyses of military pilots do however reveal slight differences in the requirements in comparison to civilian airline pilots. This leads to distinct priorities in the applied selection tests. A review of identified outcome measures for the selection of military and civilian pilots was provided by Laws (2018), representing the Aerospace Medicine Systematic Review Group (AMSRG).

A good summary of the current status in pilot selection was provided by the Federal Aviation Administration in 2019. Broach, Schroeder & Gildea (2019) recommend a set of seven features for pilot selection: 1) conduct a job analysis; 2) define measurable job performance metrics; 3) identify and use reliable and valid predictors; 4) conduct an appropriate validation study; 5) determine cut-scores on tests based on predicted job performance; 6) evaluate the fairness of tests and cut-scores; and 7) document the analyses. However, the reality showed that only a few of the 15 reviewed pilot selection batteries actually fulfil these criteria. The authors are demanding more fine-grained pilot performance measures that go beyond simple pass/fail-criteria in training as well as more complete and open documentation of pilot selection systems especially for civilian pilots.

9.2. Common Denominators for the Selection of ATCOs and Pilots

The seven FAA criteria for a scientifically solid selection system describe fundamentally good practices and do apply for both pilot selection and for ATCO selection. Another common ground of the selection procedures for both professions is that usually multiple diagnostic instruments should be used: basic aptitude tests, basic work sample tests, questionnaires (biographical, personality), behaviour-based diagnostics (e.g., assessment centre), and interviews (EUROCONTROL, 2002; IATA, 2019). Further examples are provided in chapter 6. This indeed is good practice because each diagnostic instrument can compensate (within limits) the flaws of the other instruments applied

Both EUROCONTROL and IATA have previously conducted industry surveys in Europe and beyond to collect information about the current practices in ATCO and pilot selection. Commonly used aptitudes and competencies that were included in selection systems have been compiled and presented in Table 9. 1 to Table 9. 4. The criteria listed below are not necessarily based on analyses of relevant job requirements and therefore not legally binding. They reflect criteria which are considered as selection standards for the respective positions by the majority of surveyed organizations (ANSPs and airlines) in the terminology of EUROCONTROL and IATA.

In comparison of the two profiles a large overlap can be stated especially for the basic aptitudes in Table 9. 1 and the social competencies in Table 9. 4. However, this does not imply that the same weight would be assigned to these criteria. On the other side, some of the differences are that IATA would not consider mental arithmetic and mathematical knowledge as selection standards for pilots. IATA might regard this aspect as covered by educational entrance requirements. In the area of operational competencies there is some overlap and some discrepancy (Table 9. 3). The overlap is with respect to planning, problem solving, and stress management. However, IATA criteria cover in addition two competencies suggested by ICAO:

flight path management and rule compliance. It makes sense that flight path management is required only for pilots. IATA, however, does not explain how these criteria should be measured with ab-initio candidates who are not yet familiar with the cockpit environment.

The largest differences can be found with respect to personality as shown in Table 9. 2. While EUROCONTROL found that most ANSPs selection systems cover general personality traits, IATA missed identifying these in their survey with worldwide airlines. IATA is rather suggesting that the majority of surveyed airlines are interested in attitudes towards self and a general attitude towards safety as selection criteria for pilots.

Table 9. 1: Comparison of selection criteria for ATCOs and pilots as suggested by EUROCONTROL and IATA (abilities)

Cognitive Abilities	ATCO	Pilots
Basic aptitudes	memory functions	memory capacity
	perceptual speed and accuracy for visual or auditory information	speed and accuracy of information processing (perception, classification, transformation)
	spatial comprehension: spatial orientation and visualization	spatial abilities (static)
		spatial abilities (dynamic)
	logical reasoning with verbal numerical and figural material	reasoning (information processing with basic figures)
		logic abilities
	vigilance	long-term concentration
	attention, concentration, divided attention and selective attention	allocation of attention
	mechanical and technical understanding	technical comprehension

Cognitive Abilities	ATCO	Pilots
Multi-tasking and psychomotor abilities	multiple tasks abilities	multi-tasking (different tasks combined)
	perceptual / psychomotor abilities such as motor coordination, arm-hand steadiness, reaction time tests or electrical contacts register tests	psychomotor abilities (pursuit tracking, compensatory tracking)
Knowledge	mental arithmetic	-
	mathematical knowledge	
	English language proficiency including grammar, vocabulary and syntax	English language proficiency

Table 9. 2: Comparison of selection criteria for ATCOs and pilots as suggested by EUROCONTROL and IATA (personality)

Personality Traits	ATCO	Pilots
Motivation	vocational motivation / interests	professional motivation
	need to achieve, persistence, resilience, vitality, readiness to acquire new knowledge and skills, responsibility	
Decisiona making	flexibility, creativity and dominance related to decision making behaviour	-

Personality Traits	ATCO	Pilots
Social behaviour	extroversion vs. introversion, dominance or assertiveness, empathy, aggression in relation to social behaviours	-
Stress-coping	emotional stability, flexibility and aggression in relation to stress-coping	stress-coping with social confrontation, information load, time pressure
Attitudes towards self	-	self-discipline
		self-criticism
		self-organization
General attitude		safety motivation

Table 9. 3: Comparison of selection criteria for ATCOs and pilots as suggested by EUROCONTROL and IATA (operational competencies)

Operational competencies	ATCO	Pilots
Planning and problem solving	leadership, planning, decission making	problem solving and decission making
	planning	situation awareness and management of information
Flight path management	-	aeroplane flight path management, manual control
		aeroplane flight path management, automation

Operational competencies	ATCO	Pilots
Rule compliance	-	application of procedures and compliance with regulations
Stress management	stress handling	workload management

Table 9. 4: Comparison of selection criteria for ATCOs and pilots as suggested by EUROCONTROL and IATA (social competencies)

Social competencies	ATCO	Pilots
Interpersonal	communication	communication
	cooperation/teamwork	leadership & teamwork

9.3. EAAP Practices in Selection Survey

To help answer questions regarding the current practices in selection, the EAAP Selection Working Group put together a questionnaire to gather information on selection tools currently in use by EAAP members, how cutoffs are set, about the length and complexity of selection, possibilities for retesting and how decisions are reached. Altogether 83 subject matter experts responded to this survey in 2020. A total of 38 respondents were primarily involved in selection of civilian (ab-initio) pilots and 12 respondents in the selection of civilian air traffic controllers. The remaining 33 participants referred their ratings to both professions simultaneously or to other relevant occupations (e.g., cabin attendants or military personnel). The majority of the criteria recommended by EUROCONTROL and IATA seem to be included in today's test batteries.

While interpersonal skills (except Leadership) received special emphasis (average ≥ 90%), basic knowledge (except English) areas (average ≤ 33%) were slightly downplayed. With respect to the criteria which form part of the respective selection process some key aspects seem to discriminate between the profiles for ATCOs and pilots. Areas where the percentages between ATCOs and pilots differed over 30% are shown in Table 9.5

Table 9. 5: Discriminating attributes between ATCO and pilot selection batteries

Attribute	% of ATCO Selection	% of Pilot Selection
Psychomotor coordination	8%	63%
Manual control	0%	58%
Mathematical knowledge	25%	55%
Aeronautical knowledge	0%	55%
Leadership	8%	92%

9.4. Summary

The current practices in ATCO and pilot selection as observed by EAAP in its recent survey seem to be in line with recommendations by EUROCONTROL and IATA. The relevant selection criteria for both professions can be assigned to the categories of cognitive abilities, personality, operational competencies, and social competencies. However, the wording for the different aptitudes and competencies differs quite significantly. It is also not clear how exactly these important aptitudes and competencies can best be measured. Therefore, it is inevitable that each organization using selection tools to recruit its personnel documents in detail which tests and measures they use, what their psychometric properties are and how the measures are weighted in relation to an in-house job analysis.

9.5. References

American Psychological Association (APA) (2018). Principles for the validation and use of personnel selection procedures (5th ed). Society for Industrial and Organizational Psychology, August, 2018.

American Psychological Association (APA) (2017). Ethical principles of psychologists and code of conduct: Section 9 Assessment. Retrieved from https://www.apa.org/ethics/code/

Broach, D., Schroeder, D. & Gildea, K. (2019). Best practices in pilot selection. FAA Final report DOT/FAA/AM-19/6. Federal Aviation Administration, Office of Aerospace Medicine, Washington, DC 20591.

EFPA Board of Assessment (2013). EFPA review model for the description and evaluation of psychological and educational tests, Document 110c version 4.2.6. EFPA: Brussels.

European Commission (EC) (2018). Commission regulation (EU) 2018/1042 of 23 July 2018. Official Journal of the European Union, L 188/3. 25.07.2018. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1539167289501&uri=CELEX:32018R1042 .

European Union Aviation Safety Agency (EASA) (2019). Acceptable Means of Compliance (AMC) and Guidance Material (GM) to Annex IV Commercial air transport operations (Part-CAT). Retrieved from https://www.easa.europa.eu/document-library/acceptable-means-of-compliance-and-guidance-materials .

EUROCONTROL (2000). Marketing and communication of the ATCO job - Phase 1-Study. Document identifier HRS/MSP-002-REP-02. EUROCONTROL Headquarters, Brussels.

EUROCONTROL (2001a). Guidelines for selection procedures and tests for ab initio trainee controllers (Revised). European Air Traffic Management Programme. Document identifier HRS/MSP-002-GUI-01. EUROCONTROL Headquarters, Brussels.

EUROCONTROL (2001b). Characteristics of recruitment and pre-selection of ab initio trainee controllers (Revised). European Air Traffic Management Programme. Document identifier HRS/MSP-002-GUI-02. EUROCONTROL Headquarters, Brussels.

EUROCONTROL (2002). Selection tests, interviews and assessment centres for ab initio trainee controllers: Guidelines for implementation (Revised). European Air Traffic Management Programme. Document identifier HRS/MSP-002-GUI-03-01. EUROCONTROL Headquarters, Brussels.

EUROCONTROL (2003). Marketing and communication of the ATCO job - Phase 2-Study: The ATCO job as perceived by its applicants. Document identifier HRS/MSP-002-REP-03. EUROCONTROL Headquarters, Brussels.

International Air Transport Association (IATA) (2019). Pilot aptitude testing (Edition 3) – Guidance and best practices. IATA: Montreal-Geneva.

International Civil Aviation Organization (ICAO) (2020). Amendment No. 7 to the Procedures for Air Navigation Services – Training (Doc 9868). State Letter AN 12/48-20/33, June 2020.

International Test Commission (ITC) (2013). ITC guidelines on test use (Final Version 1.2). Document reference: ITC-G-TU-20131008.

Laws, J. (2018). Common psychological outcome measures for the selection of civilian and military pilots. AMSRG tool: Aerospace Medicine Systematic Review Group/UK. April 2018.

10. Potential Problems & Pitfalls

Jennifer Eaglestone

In this chapter we would like to address some additional issues in the selection of aviation personnel. Several of these topics are often an afterthought or sometimes even overlooked completely. However, each one should be taken into account when setting up and running a (recruitment and) selection procedure. Although some of these topics have not been researched to a great extent we have, in many cases, made recommendations based on our experience and knowledge of best practice.

10.1. Legal Issues

Laws and guidelines pertaining to selection differ per country. Some countries have very few statutes or guidelines; others have numerous laws at international, national and local levels, as well as important professional guidelines. An aviation psychologist involved in selection therefore must ensure that any selection system (and the instruments developed for and used in the selection system) meets all pertinent national and local laws and guidelines.

Some countries, for instance, have specific laws about when certain types of testing can occur. Perhaps the best examples of this are the medical and security/background checks. At the time of writing, in the US, UK and the Netherlands for example, security and medical checks can only occur after a candidate has received a conditional offer of employment.

More general laws and directives can also be of importance. For instance, in the EU, discrimination on any ground is prohibited, be it gender, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation. This can make it very difficult to for instance put filters for selection in place such as age restrictions even if there is empirical evidence supporting it from a scientific or technical point of view. Some countries in the EU do apply age restrictions in selection (e.g. see ATCO selection in Germany and the Netherlands) but solid substantiation should be provided in case of complaints or challenges (for more information on age restrictions see paragraph 9.2).

Selection decisions should always be based on candidate-appropriate norms but test norms themselves can also be regulated by law. Some countries permit gender and age-based test norms for scoring, whereas others, like the US, do not (in the US all tests must be evaluated using the same scoring key).

Besides national and international law, there are also aviation specific regulatory guidelines that must be adhered to. For instance, as mentioned in chapter 9, the EASA Commission Regulation 2018/1042 EASA regulations on the psychological assessment of pilots (EASA 2018, EASA 2019).

10.2. Age Restrictions for Applicants

10.2.1. Pilots

Age restrictions as they pertain to civilian pilots are reasonably clear. ICAO member states set minimum age requirements for specific certificates. This has little effect on selection for ab initio programs because most air carriers with ab initio programs have educational requirements that effectively ensure that a candidate is close to the minimum age required for a student license.

ICAO Amendment 172 to Annex 1--Personnel Licensing (Woods & Ahmed, 2014) allows pilots flying multi-crew aircraft to fly until age 65. Single-pilot air transport operations are limited to age 60, although member states may impose a lower limit. In some countries, such as the US, pilots have no age restriction if they are flying cargo for a non-scheduled air carrier.

How does this age limit affect pilot selection? Setting an upper age limit for candidates (when allowed by law) is essentially a management decision. During periods of a pilot shortage, management may choose to hire experienced pilots who are close to retirement age to ensure sufficient staffing. The psychologist's role under such conditions is to ensure that selection is conducted in a non-discriminatory manner, however despite some compensation by experience a certin level of mental abilities must be ensured.

10.2.2. ATCOs

In many countries, Air Navigation Service Providers utilize age limits for ab initio applicants. In countries where national law allows for age restrictions in selection (as it does in many parts of the EU for instance), several factors can be taken into account. For lower age limits one may consider, for instance, the minimum age for a student controller license. For upper age limits several other factors must be considered, for example factors that increase or decrease with age and affect training failures and economical factors. We must not forget, however, that age limits for experienced controllers should probably be different than for ab initio's, as experience gained with age can compensate for other attributes (Broach & Schroeder 2005).

Retirement age for ATCOs is nation specific, with retirement from operations varying between 50 and 65. On occasion it seems that retirement age is taken into consideration due to economical reasons which may justify age restrictions. when setting an age limit for applicants.

It remains to be said that this is an incredibly complex subject which we only touch on here. There are many things that should be taken into consideration, many reasons for setting age limits. If dealing with this matter we also recommend you seek further advice on both a legal and psychometric level.

10.3. Retesting

Organisations and test providers vary greatly in their policies on retesting. Some organisations allow no retesting, whereas others allow retesting for applicants with specific scores and/or after a specific amount of time (see EAAP questionnaire results in Appendix B.2 General survey results). Retesting is often implemented because of organisational concerns about losing 'false negative' applicants or issues concerning the 'applicant friendliness' of an organisation.

In 2007, Hausknecht et al. carried out a meta-analysis of coaching and practice effects for tests of cognitive ability. Practice effects were identified across measurements and scores improved when applicants were retested. They do conclude, however, that a retest after more than a year should ensure that effects are minimal. Scharfen et al (2018) also showed significant retest effects that, however, subsided after the third administration of the test. Their moderator analysis indicated that cognitive ability, equivalence of test forms, retest interval and age all have a significant influence on the size of the retest effect. Albers and Höft (2007) saw performance reach an asymptote after a fifth retest. When deciding if candidates can retake certain selection rounds, it is important to be aware of the training effects.

Several researchers have posed the question if a score on the original test or the score on a retest is more representative for a person's latent ability (Lievens et al., 2005; te Nijenhuis et al., 2007), as this can lead to decision errors in all kinds of contexts (Randall & Villado, 2017). Lievens et al (2005) showed that the same test score led to higher levels of performance for those passing on the first attempt than for those passing on the second attempt, indicating that score from the first attempt is a more accurate predictor. That said from a practical point of view, from experience we can say that many organisations are suffering from a shrinking applicant pool which may lead to the decision to offer to possibility to retest after a certain amount of time. In this fashion false negatives from these rounds of testing could possibly be intercepted.

10.4. Test Practice & Familiarization

A related topic is that of test practice and familiarization. Test practice can be defined as pay-to-practice training offered by third parties. Test familiarization, on the other hand, is a service offered by the test provider or selecting organisation to enable candidates to acquaint themselves with the (type of) tests used.

The question at hand is should candidates be allowed, or even encouraged, to practice tests (actual or similar) prior to the selection process? It is considered good practice to be transparent about as many aspects of your selection procedure as possible and this can include test familiarization. Some organisations using tests without a familiarization feature guide applicants towards websites where they can practice other cognitive ability tests to become familiar with computer-based testing. One could also state the better the familiarization the less chance an applicant will pay to practice.

In many cases pay-to-practice websites or coaching on an entire selection procedure is offered by third parties. This causes both ethical and validity issues, as complete fairness cannot be guaranteed when there are differences between candidates' experiences with the tests (Albers and Höft, 2007). Candidates are also given the impression that they need to spend (sometimes significant) amounts of money to prepare for selection, whereas this should not be the case. We cannot, however, control what businesses are doing, but only warn candidates about such practices. The solution for this matter is to build / use tests that are not affected by prior practice or to offer practice tests ourselves and allow candidates to take the practice tests for free a certain number of times.

10.5. Providing Candidates with Feedback

Even though it is time-consuming, feedback is important for both the candidate (who has put time and effort into applying) and the organisation. It is not only ethically sound to offer voluntary candidate feedback, but it can also benefit the organisation. It can influence the opinion candidates have of your organisation and the extent to which they recommend it to others.

Providing candidates with feedback may be considered problematic because information on tests and the selection procedure may become public. However, providing someone with, for instance, a general profile based on their results will not provide too much information on the tests and the selection procedure, whereas it can satisfy candidates' curiosity about their performance.

Finally, we would like to encourage you to provide clear and timely information to applicants on how and where feedback is possible. This can improve transparency of the selection procedure and candidate satisfaction.

10.6. Organisational Culture

Organisational culture can be defined as 'a system of ideas and concepts, customs, traditions, procedures and habits for functioning in a specific macro culture' (Harris and Moran, 1981, pp 103-104). Every organisation has its own culture, which can explicitly or implicitly affect the behaviour of individuals. It, therefore, should not be overlooked when selecting people for an organisation (on the other hand the organisational factors must also not be overrated compared to the actual task requirements).

This in turn, means that although profiles for aviation personnel will have a great deal of overlap, they are not generic. Apart from the attributes operational personnel may need for the task at hand (for instance the general profile for an ATCO), there may also be culture-specific attributes that differ between organisations (e.g. cargo pilot vs passenger airline pilot; military vs civilian, European legacy airline vs more multicultural Middle Eastern airlines). This underlines the importance of organisation-specific job analysis (for more information on job analysis chapter 5).

10.7. Job Performance Data

According to Schmidt & Hunter (1992), job performance is the most important dependent variable in Industrial Psychology. Job performance data plays a crucial role in recruitment and selection, from determining which source of recruitment (e.g. TV, newspaper ads, LinkedIn) attracts the best candidates to validating selection instruments (Viswesvaran & Ones, 2017).

Every selection system (see chapter 7) should be validated at regular intervals and ideally validated using job performance data (chapter 5, chapter 8). Predictive validity can decrease over time for a number of reasons. For example, selection instruments may be compromised, or the applicant population may undergo a subtle change. Changes in both training and the job itself (for instance increasing automation) may also affect the predictive validity of the selection system.

Obtaining reliable data for validation is often more difficult in aviation than in some other fields. Training data are typically reasonably easy to obtain but job performance data are more difficult to access. Data may be protected by data storage regulations (see 10.8), a union agreement making them inaccessible or even just difficult to obtain because the operators themselves are opposed to it. Performance assessment may suffer from ceiling effects and in many cases methods do not allow for discrimination between individuals' performance (and there is often only pass/fail data recorded). Accident, incidents, and safety violations are rare events in aviation, making them inappropriate for validation. These problems and others pose serious challenges to any validation effort.

We encourage you to set up performance datapoints (set moments in time where data is collected about an individual's performance) within your organisation to enable you to validate your selection (Broach et al. 2019) These can be comprised of both training data and specific job performance data collected for selection validation. For more general information on collecting job performance data see Viswesvaran & Ones (2017).

10.8. Data Storage

The way we store both selection and job performance data are often protected by law. In the European Union for instance the EU General Data Protection Regulation (2016) dictates how private data should be stored, how long it may be stored and how it can be de-identified (made anonymous). For validation purposes you need both selection and job performance data, but the data needs to be linked at an individual level to be useful. As well as adhering to common sense and (inter)national data protection laws, you should also store your data in such a fashion that it can be used for your validation studies. It is advisable to have carefully defined rationale and make sure you have received advice on data storage from a data protection law expert.

10.9. Online Testing

Unsupervised online testing, or unproctored internet-based testing (UIT) as it is often referred to in the literature, has now become a normal part of many selection procedures. This type of testing is carried out online wherever and whenever the candidate finds it most appropriate. Although not suitable for all types of psychometric testing (e.g. psychomotor abilities), in a 2015 global survey, 40% of organisations surveyed reported that they used UIT for all of their test procedures (Ryan et al. 2015). These tests offer a low-cost way of reaching large groups all over the world.

They do, however, come with their own set of challenges. Woods et al. (2020) give an excellent review of the literature on online testing concluding that the implementation of UIT is widespread, while scientific research is still somewhat lacking. Some organisations are using online testing without sufficient knowledge of the validity of the tests and the impact on applicants.

One of the main issues with UIT is cheating. Although estimates for the percentage of candidates that try to cheat are broad, ranging from 7 – 50% (Arthur et al. 2010), effective cheating will always have a significant impact on selection and damages the fairness of the procedure (Bloemers et al. 2016). Some researchers feel that administrators would be well advised to check for cheating post hoc (e.g. Steger et al. 2020). Others advise to "construct and use unproctored test batteries consisting of items with a high g load (complex reasoning)" (Bloemers et al. 2016, p.26). Still others, such as Karim et al. (2014), consider remote proctoring (i.e. supervision via a webcam) to be the best way of combatting fraud.

10.10. Summary

In this chapter we have addressed several topics that must be considered when constructing a selection procedure. These range from (inter)national law and regulations to more ethical issues, such as feedback and cheating. These topics should not be overlooked and will help you in making certain decisions for your selection procedure.

10.11. References

Arthur, W. Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet based tests of cognitive ability and personality. International Journal of Selection and Assessment, 18, 1-16.

Albers, F. & Höft, S. (2007). Practice effects on test-takers performance and quality of cognitive ability tests in pilot selection. 14th International Symposium on Aviation Psychology, Wright State University, Dayton, OH.

Broach, D. & Schroeder, D.J. (2005). Relationship of Air Traffic Control Specialist Age to En Route Operational Errors. (DOT/FAA/AM-05/22) Washington, DC: Federal Aviation Administration Office of Aviation Medicine.

Broach, D., Schroeder, D. & Gildea, K. (2019). Best practices in pilot selection. FAA Final report DOT/FAA/AM-19/6. Federal Aviation Administration, Office of Aerospace Medicine, Washington, DC 20591.

Bloemers, W., Oud, A., & Dam, K. van (2016). Cheating on unproctored internet intelligence tests: Strategies and effects. Personnel Assessment and Decisions: Vol. 2: Iss. 1, Article 3.

European Commission (EC) (2018). Commission regulation (EU) 2018/1042 of 23 July 2018. Official Journal of the European Union, L 188/3. 25.07.2018. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1539167289501&uri=CELEX:32018R1042 .

European Union Aviation Safety Agency (2019). Acceptable Means of Compliance (AMC) and Guidance Material (GM) to AnnexIV Commercial air transport operations [Part-CAT] of Commission Regulation (EU) 965/2012 on air operations. Retrieved from https://www.easa.europa.eu/sites/default/files/dfu/Consolidated%20AMC-GM_Annex%20IV%20Part-CAT_March%202019.pdf

EU General Data Protection Regulation (GDPR). 2016.: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Handy C. B. (1986). Understanding Organizations, 3rd ed. Penguin Books, Harmondsworth

Harris P.R., R. T. Moran (1981). Managing Cultural, 2nd ed., Gulf Publ.Co. Huston

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. Journal of Applied Psychology, 92(2), 373–385

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. Journal of Business and Psychology, 29, 555–572.

Lievens, F., Buyse, T. & Sackett, P.R., (2005). Retest effects in operational selection settings: Development and test of a framework. Personnel Psychology, 58: 981-1007.

Nijenhuis, J. te., Vianen, A.E.M. van., & Flier, H. van der. (2007). Score gains on g-loaded tests: No g.Intelligence, 35-3, 283-300, https://doi.org/10.1016/j.intell.2006.07.006.

Nikolaou, I., Costa, A.C. & Anderson, N.R. (2020) Personnel selection in the digital age: a review of validity and applicant reactions, and future research challenges, European Journal of Work and Organizational Psychology, 29:1, 64-77, DOI: 10.1080/1359432X.2019.1681401

Randall, J.G. & Villado, A.J. (2017). Take two: Sources and deterrents of score change in employment retesting, Human Resource Management Review, 27-3, 536-553, https://doi.org/10.1016/j.hrmr.2016.10.002.

Raz, N. (2000). Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In F. I. M. Craik & T. A. Salthouse (Eds.), The handbook of aging and cognition (p. 1–90). Lawrence Erlbaum Associates Publishers.

Ryan A. M., Inceoglu I., Bartram D., Golubovich J., Grand J., Reeder M., & Yao X. (2015). Trends in testing: Highlights of a global survey. In I. Nikolaou & J. K. Oostrom (Eds.), Employee recruitment, selection, and assessment: Contemporary issues for theory and practice (pp. 136–153). Sussex, UK: Psychology Press.

Sackett, P.R. & Yang, H. (2000). Correction for range restriction: An expanded typology. Journal of Applied Psychology, 85(1), 112-118. DOI: 10.1037//0021-9010.85.1.112

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. Intelligence, 67, 44–66. http://dx.doi.org/10.1016/j.intell.2018.01.003

Schmitt, F.L. & Hunter, J.E. (1992). Causal modeling of process determining job performance. Current Directions in Psychological Science, 1, 89-92

Viswesvaran, C. & Ones, D. S. (2017), 'Job performance: Assessment issues in personnel selection.' In A. Evers, N. Anderson & O. Voskuijl (Eds.) The Blackwell handbook of personnel selection: 354-75 https://doi.org/10.1002/9781405164221.ch16

Woods, S.A. & Ahmed, S. (2014) ICAO Amendment 172 to Annex 1—Personnel Licensing.

11. Emerging Technologies

Karien Stadler

We have entered a very exciting era when considering the assessment of human behaviour, skills and abilities. Alongside the rapid technological advances in the aviation industry, we have also witnessed technical advances in selection methods and instruments. Modern-day assessments are increasingly reflecting technological trends and offer rich data sets which do not only include performance data, but also test-taking behaviour such as the number of mouse clicks and interactions, timing, audio and video. It therefore provides the possibility to realise designs that could not be appropriately implemented with the more traditional paper-pencil tests (Greiff, Scherer & Kirschner, 2017; Tippins, 2015). Although some of the current and future assessments may offer innovative breakthroughs in the way we assess people, they present many challenges such as potential bias in personality profiling.

Even though computerisation enhances standardisation, test conditions may vary due to distractions and interruptions such as internet connectivity issues. These test conditions can have a negative impact on ability test results in a selection context. However, companies are moving ahead regardless to incorporate these cost-effective technologies into their testing processes which further widens the gap between practice and the evidence base (Woods et al., 2019). Paradoxically, while the interest in digital selection procedures increases, published research is limited. Continuous research in digital selection procedures should therefore be encouraged to align theory and practice and address concerns.

The recent study by Tippins, Oswald & McPhail (2021), explore these concerns comprehensively. They recommend a useful three-part framework for understanding these new technologically enhanced forms of assessment based on different technologies, types of data, and algorithms. They also make an urgent call to industrial and organisational psychologists to extend existing professional standards for employment testing to these new forms of testing, including standards and requirements for their documentation. This chapter aims to give an overview of the trends in psychological assessments. It is by no means a comprehensive science-based review but sensitises the practitioner on current and emerging trends.

11.1. Artificial Intelligence

When thinking about the future of assessment, a buzz word is artificial intelligence. The term artificial intelligence (AI) describes the intelligent behaviour of machines that mimic the decisions, processes, or outcomes of humans (Tippins, Oswald & McPhail, 2021). It can be in the form of hardware or software, stand-alone, distributed across computer networks or embodied into a robot. It can even present in the form of an autonomous interactive virtual human that could conduct interviews for selection.

Advances in AI enables speedy processing, lower costs, convenient access and applicant engagement (Tippins, Oswald & McPhail, 2021). Proponents of this technology claims that it avoids human bias, such as unconscious preference and can therefore be more predictive of job performance than selection processes facilitated by human selection specialists. It can also evaluate large pools of candidates (Poli, 2019).

Critics however feel that AI amplifies bias and that algorithms could drive its preference toward a single type of candidate (Heilweil, 2019; Yankov, et al., 2020). Tippins, Oswald & McPhail (2021) postulate that these tools raise serious concerns about their effectiveness in terms their conceptual relevance to the job; their basis in a job analysis to ensure job relevancy; their measurement characteristics (reliability and stability); their validity in predicting employee-relevant outcomes; their evidence and normative information being updated appropriately; and the associated ethical concerns around what information is being represented to employers and told to job candidates.

11.2. Machine Learning

Machine learning (ML) is an aspect of AI that refers to the capability of computers to learn without being programmed. The computer learns from the past experiences (input data) and makes future predictions. Typical application in the selection industry includes statistical learning methods, predictive modelling, data mining, image recognition and natural language processing (Luxton, 2016). These methods bring a computational approach for working with uncertainties and can assist psychologists and HR professionals to optimise selection decisions.

For applications of ML in personnel selection settings, Tippins, Oswald & McPhail (2021) claim that, generally, the effectiveness of ML prediction or clustering will more likely be driven by availability of high-quality data as opposed to the specific ML algorithm that is chosen. Sajjadiani, et. al., (2019) researched the application of machine learning to translate applicant work history into predictors of performance and turnover. They quantified the extent to which their model can improve the quality of selection processes above the conventional methods of assessing work history, while lowering the risk of adverse impact. It should be noted that accurate and well justified predictions depend on good measurement processes as well as good data. It is in the interest of all selection practitioners to know more about ML algorithms and algorithmic bias. Even though it might not form part of your selection processes, it can be useful in conversations with test publishers and other vendors of ML algorithms (for an example of guidelines see Yankov, et al., 2020). In recent years there have been various AI regulatory and legislative developments.

The European Commission unveiled a new proposal for an EU regulatory framework on artificial intelligence in April 2021. The United States of America has also introduced new legislation in this regard, such as the Algorithmic Accountability Act of 2022 and the Artificial Intelligence Video Interview Act of 2019. These and other emerging regulatory and legislative developments should be top of mind for any practitioner or organisation using AI or algorithms.

11.3. Neural Networks

Artificial Neural Networks (ANNs) are mathematical models that have been motivated by the functioning of the brain and are at the heart of deep learning algorithms. To gather more insights on deep learning for prediction models and to understand the potential application in selection, the introductory review by Emmert-Streib, et.al., (2020) may be a good starting point.

Choosing appropriate psychometric predictors in personnel selection is an important issue in building either parametric or nonparametric classification models. Parametric classification methods (discriminant analysis, logistic regression) have been used extensively with a dichotomous pass versus fail criterion variable in personnel selection, as well as in the aviation industry, e.g., fighter pilot selection (Sommer, et.al, 2004). Over the past years there have also been efforts to build on the predictive accuracy and efficiency of nonparametric neural networks as applied to classical problems.

Maroco and Bartolo-Ribeiro (2013) evaluated the sensitivity, specificity and accuracy of traditional parametric classifiers (linear discriminant analysis, logistic regression) and four nonparametric neural networks (multilayer perceptrons, radial bias function, probabilistic neural networks and linear neural networks) devised for classification tasks in the prediction of pass versus fail pilot candidates on a flight screening programme. They found that simple, inflexible parametric classifiers might not have the power to learn the variety of interactions and direct effects underlaying the relationship between data points.

However, more complex and flexible models such as neural networks have a tendency to overfit the data and show model instability when extrapolating to new data sets, making the results nongeneralizable to a wider population. However, they propose that with small sample sizes and studies where multivariate normality and homoscedasticity of covariances can be met, linear discriminant analysis may be a simple, theoretically robust, reproducible and efficient classifier for personnel selection. Another study by Klokker, et.al., (1999) reported that a trained neural network technique could give important information contributing to meaningful selection of pilots before initiating military flight training, by excluding candidates who might not succeed.

11.4. Asynchronous Video Interviews and Natural Language Processing

Natural Language Processing (NLP) is a subfield of AI and combines linguistics, computer science and information engineering. It refers to the programming and capability of computers to process and analyse large amounts of natural language data. Recruiters and large corporate companies are increasingly using NLP in selection, for example to screen applications (Sajjadiani et al., 2019) or for asynchronous web-based video interviews (AWBVI), where interviewees record their responses to a standard set of questions that are then evaluated by either raters or computers through NLP (Guchait et al., 2014; Morelli, 2019). Verbal responses - and in some cases even facial expressions, response time and speech rate - are scored by use of an NLP program (Heilweil, 2019; Woods et al., 2019).

Recent research by Hickman, Tay & Woo (2019) raised a number of questions regarding their use. They conducted research on language-based personality assessments using an off-the-shelf, commercially available product in the context of video-based interviews. They compared scores derived from the language-based assessment to self and observer ratings of personality to examine convergent and discriminant relationships. The language-based assessment scores showed low convergence with self-ratings for openness, and with self and observer ratings for agreeableness. No validity evidence was found for extraversion and conscientiousness. For neuroticism, the patterns of correlations were in the opposite of what was theoretically expected, which raised a significant concern.

Studies on digital interviews also highlighted various concerns related to privacy, ambiguity, perceptions of fairness and social presence (Langer et al., 2017). In contrast to those findings, Suen, Chen, & Lu (2019) found no differences in perceived fairness between digital interviews that use AI or a human rater and synchronous video-interviews among job applicants. Gorman, Robinson & Gamble (2018) conducted an initial validation of asynchronous web-based video employment interviews as well as constructs frequently rated in them. They found that composite interview ratings and construct ratings of mental capability, knowledge and skills, applied social skills, and conscientiousness were significantly related to self-rated job performance. They also found that construct ratings of knowledge and skills and applied social skills were significantly associated with self-reported organisational tenure. With the growing need to efficiently and cost-effectively recruit and select from among an increasingly remote global workforce, selection practitioners should consider the potential benefits as well as risks related to AWBVI and NLP before considering it as screening methods or substitutes for face-to-face interviews.

11.5. Intelligent Machine Perception & Sensing

Machine perception and sensing is another form of AI that refers to the capability of a computer system to recognise and interpret images, sounds, smell and touch, in a manner that is similar to the way humans use their senses to relate to the world around them. These machine responses are enabled through attached hardware and software. Examples include visual sensing in drones, biosensors to detect heart activity and eye track technology to measure pilot attention and scanning ability (Luxton, 2016). Face scanning in selection has been a very controversial topic in news media. The Washington Post reviewed the pros and cons of AI applications of face scanning in hiring systems and a popular vendor recently eliminated the evaluation of facial characteristics from their interview models because it deemed that its incremental prediction was insufficient. Tippins, Oswald & MchPail, (2021) postulate that these recent complaints, as well as many of the claims made by test publishers about technology enhanced assessments, should be carefully investigated, verified, and vigorously debated.

11.6. Affective Computing

Affective computing refers to the emotion recognition by machines. It includes emotion modelling, affective user modelling and the expression of emotions by robots or virtual agents. Affective computing technology can detect, classify and respond to the user's emotions and other stimuli (Luxton, 2016). It can be effective in detecting psychological distress in high stakes selection contexts.

11.7. Virtual & Augmented Reality

Virtual reality refers to computer-generated simulated environments that enable humans to interact with virtual humans or other virtual forms of life in virtual environments. Augmented reality enhances the real physical world by super-imposing computer-generated visual graphics, sound or other stimuli with real world imagery (Luxton, 2016). The use of flight simulators is an example of immersive technology that aviation has implemented with great success. More recently, a few test publishers have gone to market with immersive assessment solutions such as situational judgement tests and realistic job previews. Since these tests are perceived more favourably by applicants, the same research dedicated to text-based SJTs is needed to guide development of these newer forms of low-fidelity simulation (Woods et al., 2019).

11.8. Gamified Assessments

Gamified assessments refer to traditional scientific psychometric tests which have some gaming components built around them to create more engaging and visually attractive tests. Gamified assessments can include elements such as points and badges which candidates earn as they progress through levels. The abilities and soft skills they are designed to measure are no different from traditional psychometric tests and measures. Research shows that gamified assessment methods can be an accurate and attractive selection method (Georgiou, Gouras & Nikolaou, 2019).

Video games can attract and engage candidates for certain jobs and organisations, which may be especially useful in a tight labour market where employers compete for talent (Tippins, Oswald & McPhail, 2021). The incorporation of game elements in the selection process might reduce faking, since desirable behaviours may be less obvious while playing a game. It could therefore improve the quality of information about applicants and prediction of job performance (Armstrong et al., 2016). It can also increase applicants' perceptions of process satisfaction and in turn, perceptions of fairness (Georgiou, Gouras & Nikolaou, 2020).

Future research on the validation and fairness of serious games and game-like assessments is necessary for the successful progression of its application (Woods et al., 2019).

11.9. Social Media

The practice of using social media and network websites (SNWs) to draw inferences about candidates' KSAOs in support of selection decisions is widespread (Roulin & Levashina, 2018; Woods et al., 2019). It has been suggested that most people expect and seem to be comfortable with employers checking their social media profile, especially in professionally-oriented SNWs, such as LinkedIn (Chamorro-Premuzic et al., 2016). Aguado et al., (2019) conducted an exploratory factor analyses on coded features of LinkedIn profiles.

They reported that these factors were predictive of productivity and work outcomes. Roulin & Levashina, (2018) examined the properties of LinkedIn-based assessments in two studies and found that raters reach acceptable levels of consistency in their assessments of applicant skills, personality and cognitive ability. However, many authors caution over the usage of social media platforms in employee recruitment and selection due to various issues such as weak correlations for cognitive ability and personality constructs, applicant reactions, impression management and adverse impact (Woods et al., 2019). It is probably also less suitable for ab -initio candidates who have short resumes.

11.10. Summary

This summary is by no means exhaustive but provides a basic overview of the most prominent current and future developments in the assessment field. When considering the pace of technological advances and the utility thereof in the assessment industry, aviation psychologists and selection specialists need to be agile to keep up with industry trends. In order to enhance the candidate experience, a modern selection process should resonate with the new generation candidate who has been born and raised in the digital age with mobile devices, immersive video games and the internet of things. It is therefore envisaged that these new types of assessments may become more popular in the years to come.

That being said, the review of scientific evidence highlights many limitations and amplifies the need to conduct research at a pace that reflects the speed of the emerging technological advancements. Although validity studies and other research conducted by vendors and assessment suppliers are welcomed, it should ideally be conducted by a third parties or be subjected to a peer review process to ensure objectivity and scientific rigor. Tippins et al., (2021) explored the scientific, legal and ethical concerns about AI-based personnel selection tools and made an urgent call to industrial and organizational psychologists to extend existing professional standards for employment testing to AI and machine learning based forms of testing.

In conclusion, as with all selection methods, when applying new technologies, validity and reliability criteria should be met and digital selection processes should comply with professional and ethical standards.

11.11. References

Armstrong, M. B., Landers, R. N., & Collmus, A. B. (2016). Gamifying recruitment, selection, training, and performance management: Game- thinking in human resource management. In D. Davis & H. Gangadharbatla (Eds.), Handbook of research on trends in gamification (pp. 140–165). Hershey, PA: Information Science Reference. doi:10.4018/ 978-1-4666-8651-9.ch007.

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? Industrial and Organizational Psychology, 9, 621–640.

Emmert-Streib F, Yang Z, Feng H, Tripathi S and Dehmer M (2020) An Introductory Review of Deep Learning for Prediction Models With Big Data. Front. Artif. Intell. 3:4. doi: 10.3389/frai.2020.00004

Greiff, S., Scherer, R., & Kirschner, P. A. (2017). Some critical reflections on the special issue: Current innovations in computer-based assessments. Computers in Human Behavior, 76, 715–718. https://doi.org/10.1016/j.chb.2017.08.019.

Guchait, P., Ruetzler, T., Taylor, J., & Toldi, N. (2014). Video interviewing: A potential selection tool for hospitality managers–A study to under- stand applicant perspective. International Journal of Hospitality Management, 36, 90–100.

Georgiou, K., Gouras, A., & Nikolaou, I. (2019). Gamification in employee selection: The development of a gamified assessment. International Journal of Selection and Assessment, 27, 91–103.

Georgiou K, and Nikolaou I. Are applicants in favor of traditional or gamified assessment methods? Exploring applicant reactions towards a gamified selection method. Computers in Human Behaviour, March 2020.

Gorman C.A., Robinson J., & Gamble J.S. (2018). An investigation into the validity of asynchronous web-based video employment-interview rating. Consulting Psychology Journal Practice and Research 70 (2):129-146. DOI:10.1037/cpb0000102.

Harwell, D. (2019, November 6). Rights group files federal complaint against AI-hiring firm HireVue, citing "'unfair and deceptive' deceptive" practices. The Washington Post. Retrieved from https://www.washingtonpost.com/ technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-un-fair-deceptive-practices/

Heilweil, R., (2019, Dec 12). Artificial intelligence will help determine if you get your next job. Retreived from https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen.

Hickman, L., Tay, L., & Woo, S. E. (2019). Validity evidence for off-the-shelf language-based personality assessment using video interviews: Convergent and discriminant relationships with self and observer ratings. Personnel Assessment and Decisions, 3, 10.

Jansz, J., & Tanis, M. (2007). Appeal of playing online first person shooter games. Cyberpsychology & Behavior, 10(1), 133–136. doi:10.1089/cpb.2006.9981 PMID:17305460.

Klokker, Brock-Nannestad, Mikines, Johnsen, Garhoj, & Vesterhauge. (1999). Vestibular autorotation test in primary selection of military student pilots using neural networks. Aviation, Space and Environmental Medicine, 70(10), 959-961.

Koivisto, J., & Hamari, J. (2014). Demographical differences in perceived benefits from gamification. Computers in Human Behavior, 35, 179–188. doi:10.1016/j.chb.2014.03.007.

Langer, M., König, C. J., & Krause, K. (2017). Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings. International Journal of Selection and Assessment, 25, 371–382.

Luxton, D.D., (2016). Artificial intelligence in behavioural and mental health care. Academic Press: Elsevier.

Maroco, J., & Bartolo-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. International Journal of Aviation Psychology, 23(4), 130-152.

Morelli, N. , (2019). Artificial Intelligence in talent assessment and selection. SIOP White Paper Series.

Polli, F. (2019, October 29) Using AI to eliminate bias from hiring. Harvard Business Review. Retrieved from https://hbr.

Roulin, N., & Levashina, J. (2019). LinkedIn as a new selection method: Psychometric properties and assessment approach. Personnel Psychology, 72(2), 187-211. https://doi.org/10.1111/peps.12296

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predic- tors of performance and turnover. Journal of Applied Psychology. Advance online publication.

Sommer, M., Olbrich, A., & Arendasy, m. (2004). Improvements in personnel selection with neural nets: A pilot study in the field of aviation psychology. The International Journal of Aviation Psychology, 14, 103-115.

Suen, H. Y., Chen, M. Y. C., & Lu, S. H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? Computers in Human Behaviour, 98, 93–101.

Tippins, N. T. (2015). Technology and assessment in selection. Annual Review of Organizational Psychology and Organizational Behavior, 2, 551–582. https://doi.org/ 10.1146/annurev-orgpsych-031413-091317.

Tippins, Nancy T.; Oswald, Frederick L.; and McPhail, S Morton (2021) “Scientific, Legal, and Ethical Concerns About AI-Based Personnel Selection Tools: A Call to Action,” Personnel Assessment and Decisions: Number 7 : Iss. 2 , Article 1. DOI: https://doi.org/10.25035/pad.2021.02.001. Available at: https://scholarworks.bgsu.edu/pad/vol7/iss2/1

Washington Post. A face-scanning algorithm that increasingly decides whether you deserve the job. https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. Annual Review of Organizational Psychology and Organizational Behaviour, 2, 295–322.

Woods, S.A., Ahmed, S., Nikolaou, I., Costa, A.C. & Anderson, N.R (2019) Personnel selection in the digital age: a review of validity and applicant reactions, and future research challenges, European Journal of Work and Organizational Psychology, 29:1, 64-77, DOI: https://doi.org/10.1080/1359432X.2019.1681401.

Yankov, G. P., Wexler, B., Haidar, S., Kumar, S., Zheng, J. & Li, A. (2020). Algorithmic Justice. SIOP White Paper Series.

12. Final Remarks

Diane Damos

This report was designed to guide the reader through the creation of a selection system by presenting relevant topics in the order in which they would be considered during the development process. Consequently, the report begins with the reasons for performing selection, then provides literature reviews for pilot and ATCO selection, followed by a discussion of job analysis and the types of selection instruments available. These topics are followed by chapters on the design and validation of selection systems. Current practices and problems and pitfalls are presented next. The final chapter of this paper describes some of the new technologies that will, or are, affecting selection.

The working group structured the report in this manner to aid readers with limited experience in selection. The structure reflects the sequence of issues that a selection system developer would face while constructing the system. Much of the information presented in these chapters is common to many areas of selection, i.e., it is not specific to aviation.

Two chapters, the review of pilot selection and the review of ATCO selection, present information that is specific to aviation. Unlike many reviews, both chapters are organized chronologically to allow the reader to follow the development of specific areas of selection (motor skills, spatial ability, etc.) over time. Both chapters are lengthy, but neither is exhaustive because of the amount of research conducted on pilot and ATCO selection.

The chapter on selection methods and instruments describes methods and instruments that are used in selecting individuals for many different types of positions. Each method and instrument has a section describing specifically how it is used in selecting pilots and ATCOs. The types of instruments used by EAAP members for selection were collected in a survey; the results of the survey are presented in Appendix B. It is worth reiterating that the working group and EAAP do not endorse any commercially available product. We would also like to mention that this report has not attempted to take employment laws on tests, testing or selection into consideration, because of the vast number of national differences. Nevertheless, EAAP members should make sure to be familiar with national laws that may affect the selection process.

The methodology and statistics associated with selection are constantly evolving. Developments in technology continue to open up new areas for investigation and new ways of conducting selection. New selection tools are also appearing frequently. This report, therefore, presents the state of selection in aviation at one point in time.

EAAP members consequently should use this report as a starting point and continually monitor journals and other sources of information for new developments.

13. Appendix A: Selection Checklist

13.1. Selection Method Checklist

The purpose of this checklist is to assist selection specialists with a description of the most important technical aspects to consider about assessment methods, in order to enhance its use and inform selection decisions. This checklist is aligned with the EFPA test review model and the British Psychological Society test review information pack for test publishers/distributors[6] .

CRITERIA	DESCRIPTION	YES	NO	N/A
Technical Overview				
Classification [EFPA 3.1]	Content domains are clearly specified, e.g. ability, personality, values, motivation etc.			
Intended or main area(s) of use [EFPA 3.2]	Clearly specified areas of use, e.g., occupational, clinical etc.			
Populations for intended use [EFPA 3.3]	Description of the populations for which the test is intended, e.g., general adult population, pilots, ATCs.			
Theoretical foundations [EFPA 7.1.1]	A detailed description of the constructs measured.			
Rationale [EFPA 7.2.1]	Logical and clearly presented description of what it is designed to measure and why it was constructed as it was.			
Development [EFPA 7.2.2.1]	Accurate description of the development as well as adaptation of the measure (if applicable).			

[6] The British Psychological Society – Psychological Testing Centre. Electronic document version Feb 2018. Downloaded from https://ptc.bps.org.uk/sites/ptc.bps.org.uk/files/guidance_documents/test_review_and_information_pack_for_test_publishers_and_distributors_feb_2018.pdf

CRITERIA	DESCRIPTION	YES	NO	N/A
Translation [EFPA 7.2.2.2]	Adequate description of the translation of the measure (if applicable).			
Scales [EFPA 3.4]	The various scales and variables measured by the instrument are adequately described.			
Response format [EFPA 3.5]	Adequate description of the response format (e.g., oral, paper and pencil, online) and special procedures should be described where applicable.			
Special demands on the candidate [EFPA 3.6]	A clear description of any special demands on the candidate (e.g., handedness, command of test language etc.).			
Items format [EFPA 3.7]	An overview of the items format is provided (e.g., multiple choice, true-false, Likert-type, open-ended).			
Scale type [EFPA 3.8]	Reference to the scale type should be clearly indicated (e.g., (ipsative/ normative).			
Total number of test items [EFPA 3.9]	The total number of test items as well as the number of items per scale or sub-set should be indicated.			
Different forms [EFPA 3.13]	An indication of whether different versions (e.g. short and long version, ipsative and normative version) are available should be provided.			
Psychometric properties				
Norms [EFPA 7.2.4]	Clear and detailed information provided about sizes and sources of norm groups, representativeness, conditions of assessment etc.			

CRITERIA	DESCRIPTION	YES	NO	N/A
Reliability [EFPA 7.2.5]	Comprehensive explanation of reliability and standard error of measurement (SEM), as well as the generalisability of the assessment instrument.			
Validity [EFPA 7.2.6] [EFPA 7.2.6]	Comprehensive explanation of construct validity with a wide range of studies clearly and fairly described. Comprehensive explanation of criterion validity with a wide range of studies clearly and fairly described.			
Administration				
Method of administration [EFPA 3.11]	The mode of administration should be clearly specified (e.g., supervised/ proctored, unsupervised, computerized web-based etc.).			
Administration time [EFPA 3.12]	Indication of time required to complete the assessments (preparation, administration, scoring, analysis and feedback).			
Instructions for test administration [EFPA 7.3]	Adequate instructions for test administration are provided. This includes technical and operational requirements of the test (i.e., hardware and software, screen design conventions, including where instructional text and prompts are placed, and how instructions can be accessed once testing begins.			
Practice effects [EFPA 9.1.9]	The issue of practice effects should be dealt with adequately.			
Scoring				

CRITERIA	DESCRIPTION	YES	NO	N/A
Scoring procedure [EFPA 4.1]	Overview of the rules underlying the scoring of the assessments are provided, whether it is system-based or hand-entered into the computer.			
Score transformation [EFPA 4.4]	Description of the score transformation for standard scores is provided.			
Interpretation and reporting of scores [EFPA 7.3.4]	Detailed advice on interpreting different scores, understanding normative measures and dealing with relationships between different scales; also advice on how to deal with the possible influence of inconsistency in answering, response styles, faking, etc.			
Feedback				
Feedback guidelines [EFPA 7.3.5]	Detailed guidelines and ethical considerations when providing candidate feedback.			
Computer generated reports [EFPA 7.2.8]	Clear and detailed information provided about format, scope, reliability and validity of computer-generated reports.			

13.2. Best Practice Guidelines for a Positive Candidate Experience

The purpose of these guidelines as suggested by the SIOP White paper[7] is to bridge the gap between empirical knowledge and applied practice by outlining how organisations can ensure that their selection system is well received by all candidates (Bauer et al., 2012).

7 Bauer, T. N., McCarthy, J., Anderson, N., Truxillo, D. M., & Salgado, J. F. (2012). What we know about applicant reactions on attitudes and behavior: Research summary and best practices. Bowling Green, OH: Society for Industrial and Organizational Psychology.

Procedural justice consideration	Description	Recommendations for practice
Job relatedness	Extent to which a test appears to measure the content of the job or appears to be a valid predictor of job performance.	- Base the development of the selection system on a job analysis to identify the (KSAOs) that are relevant for the job. - Use KSAOs in the design of tests and/or interviews. - Base the selection system on scientific evidence.
Opportunity to perform	Having adequate opportunity to demonstrate one's KSAOs in the testing situation.	- Ensure the selection system is comprised of multiple components to enable a comprehensive assessment of candidate KSAOs. - Ensure that the selection process is validated over time to keep it relevant.
Consistency	Uniformity of test content, test scoring, and test administration.	- Use standardized tests and interviews based on extensive job analysis. - Provide extensive training to test/interview administrators to ensure that standard procedures are followed for all candidates. - Ensure all materials (online and elsewhere) send consistent messages regarding your organisation.

Procedural justice consideration	Description	Recommendations for practice
Feedback	Providing candidates with informative and timely feedback on the decision- making process.	- Use a computerized application system whereby candidates can track their progress and view results of the decision-making process. - Make timely feedback a priority and track time- to-feedback for each selection hurdle.
Explanations & justification	The provision of justification for a selection decision and/or procedure.	- Give candidates as much information as possible. - Put the information in context such as the number of applicants, number of vacancies etc. - Provide candidates with information regarding future job applications.
Honesty	The importance of honesty when communicating with candidates.	- Ensure that the process is transparent. - Train and reward administrators for being honest with candidates. - If providing negative results, focus on the facts and not personal characteristics. - Ensure all materials (online and elsewhere) are accurate messages regarding your organisation

Procedural justice consideration	Description	Recommendations for practice
Treat candidates with respect	The degree to which candidates feel they are treated with warmth and respect by test administrators.	- Treat candidates with respect. - Provide interpersonal training for all administrators. - Highlight the importance of ensuring that the organisation is perceived in a positive light.
Two-way communication	The interpersonal interaction between the candidate and test administrator that allows candidates the opportunity to have their views considered.	- Train interviewers to be good listeners. - Include open-ended questions as part of the standardized test process that allow candidates to ask questions.

14. Appendix B: Results from EAAP Member Survey

Hans-Jürgen Hörmann, Karien Stadler & Jóhann Wium

14.1. Introduction

In 2019 the European Association for Aviation Psychology (EAAP) formed the Working Group on Psychological Selection. The working group decided on the creation of a report on the topic of aviation selection and to survey EAAP members on how they were carrying out selection of aviation personnel (Eaglestone, Damos, Hörmann, Stadler & Wium, 2022).

The main intention of the EAAP survey was to compile the current selection practices of EAAP-affiliated practitioners working for airlines, air navigation service providers (ANSP) and other aviation organizations. Questions were asked about selection methods, instruments and other procedural aspects related to selection (e.g. decision making, norms, cut-off). The survey did not ask questions on number of applicants, vacancies, selection ratio, pass/fail rates or other similar metrics as the intent was to capture how selection was carried out (in particular for different operational groups in aviation such as pilots, air traffic controllers and cabin crew) rather than organisational specifics.

Two previous studies with a comparable survey could be located in the literature, one by the International Air Transport Association (IATA) for civilian pilots and one by EUROCONTROL for ab-initio air-traffic control trainees. IATA distributed an online survey in 2009 to 327 aircraft operators from around the world. Altogether 91 questions were asked, partly about the organizational and financial aspects of their selection process. The last part of the survey focused on the structure and contents of the pilot selection methods themselves. Only 66 of the 327 addressed organizations (20.2%) completed the survey, in whole or partially. Detailed information about the contents and conditions of the selection procedure was shared by less than 10%. It was identified that most pilot selection systems lacked a conceptual basis. Especially, selection systems for experienced first officers and captains seemed to be less sophisticated, while ab-initio systems appeared more mature. Only about 42% to 50% of the selection criteria were based on scientific analyses of job requirements. Further weaknesses reported by the organisations were (a) lacking qualification requirements for the test-operators, (b) a low "degree of automation", and (c) the time-consuming efforts involved. The strengths were seen in the high reliability and quality of the evaluation procedure. The most prevalent instruments at all levels of experience were questionnaires and semi-standardized interviews. In addition, at the ab-initio level computer-based psychometric tests of mental abilities and personality and for licensed pilots full-flight simulators checks were administered. Two-thirds of respondents preferred at least a two-step selection process. In 70% of the cases, candidates received information about the selection procedure in advance. Re-applications were possible in 60% of the organizations. Only half of the organizations involved psychologists the selection process. Despite the limited response rate, IATA based its published guidance material and best practices manual for pilot aptitude testing on results from this survey (IATA, 2012, 2019).

Information on selection tools and methods for ab-initio trainee controller selection was gathered in 1996 by EUROCONTROL's Selection Task Force (STF). The intention was to get an overview and detailed information on available or emerging tools and methods. A total of 63 institutions in 52 different nations were approached and 34 (53.9%) returned the information (EUROCONTROL, 2001).

A second wave of data collection was done between November 1999 and May 2000 as part of the development of the First European ATCO Selection Test package (FEAST). 51 providers received the survey in 35 nations and 44 (86.3%) returned the required information.

Notable figures from EUROCONTROL's surveys was that at least 83% used interviews as part of their selection process, with 52% using biographical interviews, 16% using situational interviews and 32% using mixed type of interviews. The interviews could be intended to measure general motivation (88%), specific job-oriented motivation (83%), personality (75%), communication skills (71%), teamwork skills (63%), biographical elements (63%), punctuality (50%) or stress resistance (46%) (EUROCONTROL, 2001).

A slightly higher percentage stated they used ability tests in their selection. Organisations using ability tests used them to measure: multi-tasking (76%), logical reasoning (73%), memory capacity (70%), spatial orientation (70%), mental arithmetic ability (70%), verbal English ability (67%) or written English ability (53%). While the attributes assessed therein are not mentioned, 30% of organisations used some form of assessment centers in their selection (EUROCONTROL, 2000).

Re-applications were possible in 25% of cases (EUROCONTROL, 2001) and 23% of respondents said they use mostly or exclusively paper-and-pencil testing, with 33% stating that they had moved to using mostly or exclusively computer-based testing (EUROCONTROL, 2000).

As previously stated, the intent of the EAAP survey was to determine how EAAP-affiliated professionals were selecting for different aviation roles. While the answers provide a useful indicator of distribution and frequencies, where possible we have tried to compare the results from the survey with comparable data (i.e the

IATA and EUROCONTROL surveys). On that basis the state of professional rigour in aviation selection can be inferred, developments can be estimated and current weaknesses and opportunities for advancement and growth determined.

14.2. Method

In April 2020, the EAAP Working Group for Selection in Aviation distributed an online survey to all EAAP members and some EAAP partners (N = 508) to share the expertise in selection of operational staff in aviation. The survey was open for two months.

14.2.1. Sample

A total of N = 83 participants responded to this survey. Since several of the email addresses on the distribution list were outdated, the invitation mail was bounced back and did not reach the recipients. Therefore, an estimate of the actual response rate is approximately between 16.3% and 20%. Respondents were 45% female and 53% male with an average of 17 years' experience in selection of aviation personnel. A description of all sample characteristics is shown in Table 1.

Characteristic	Value	Frequency	
		Absolute	Percent
Total sample	N	83	100%
Gender	Female	37	45%
	Male	44	53%
	Undisclosed	3	2%
Geographic region	Europe	72	87%
	Non-European	11	12%
Ocupational background	Aviation Psychologist	61	73%
	Human Factors Specialist	12	14%
	ATCO	1	1%
	Pilot	7	1%
	Other	2	2%

Characteristic	Value	Frequency	
		Absolute	Percent
Work environment	Aircraft operator (civilian)	25	30%
	ANSP	14	17%
	Authority & AMC	7	8%
	Consultancy	17	20%
	Military	5	6%
	Training facility	3	4%
	Research & university	11	13%
	Other	1	1%
Years of experience	1 to 5 years	8	10%
	5 to 10 years	12	15%
	10 to 20 years	30	36%
	over 20 years	33	40%
Selecting primarily	Civilian fixed wings pilots	43	52%
	Military fixed wings pilots	11	13%
	Civilian helicopter pilots	4	5%
	Civilian ATCOs	13	16%
	Military ATCOs	2	2%
	Cabin crew members	5	6%
	Others	5	6%

Table 1: EAAP-Survey 2020 - Sample characteristics

14.2.2. Survey

The survey intended to capture the pre-COVID conditions with altogether 27 questions. Six questions were related to demographics and the present work environment. Twenty-one questions addressed the measured selection criteria, the used selection instruments and test technologies, the data processing and decision-making aspects. In most instances a multiple-choice answer format was chosen with the opportunity to add explanatory comments.

14.3. Results

In order to compare the main selection criteria for the different positions, we decided to leave out the helicopter pilots because of the low number of answers and to categorize military and civilian ATCOs into one group. This led to a comparative analysis of four groups: 43 civilian pilots, 11 military pilots, 15 ATCOs, and 5 cabin crewmembers. The compared selection criteria were grouped into Knowledge, Skills (technical and interactive), Abilities, and Others (personality, interests, attitudes). Results are shown in Figure 1 to Figure 5 below. Because the unequal group sizes statistical significance testing was not conducted.

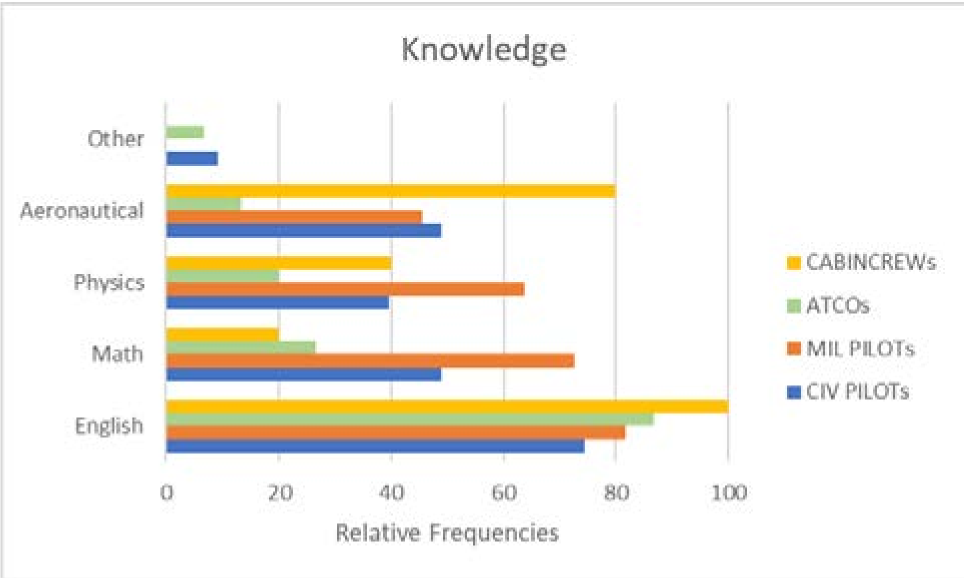

Figure 1: Relative frequencies of knowledge related selection criteria

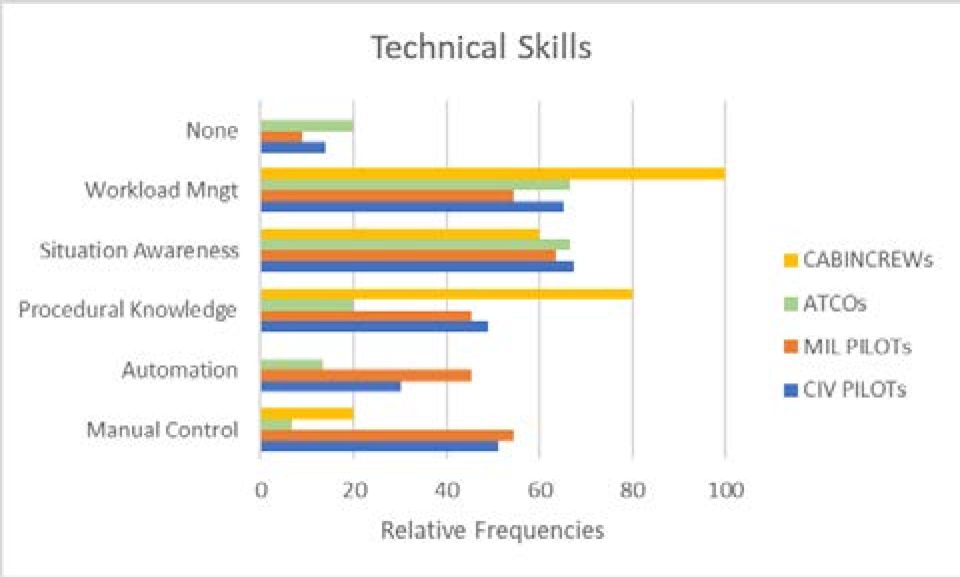

Figure 2: Relative frequencies of technical skills related selection criteria

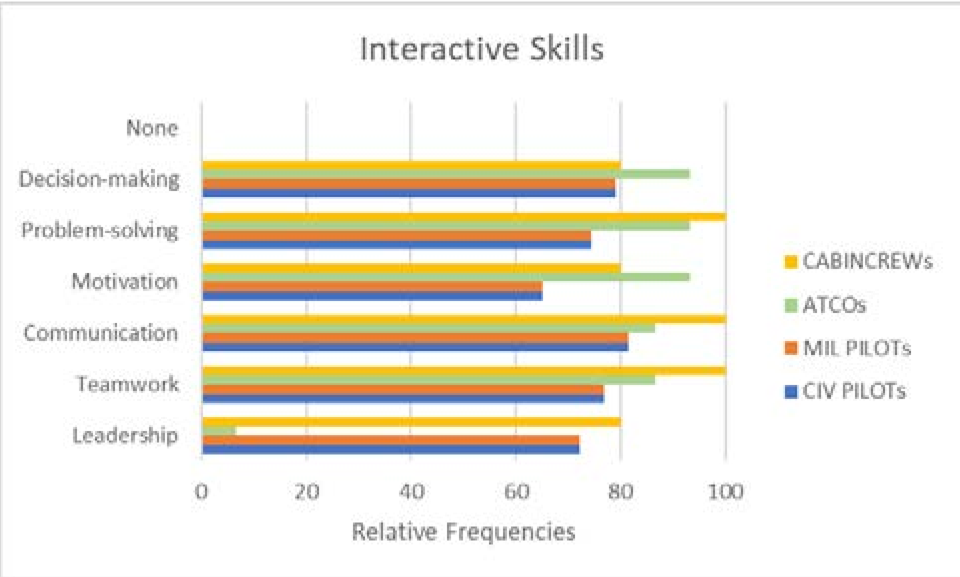

Figure 3: Relative frequencies of interactive skills related selection criteria

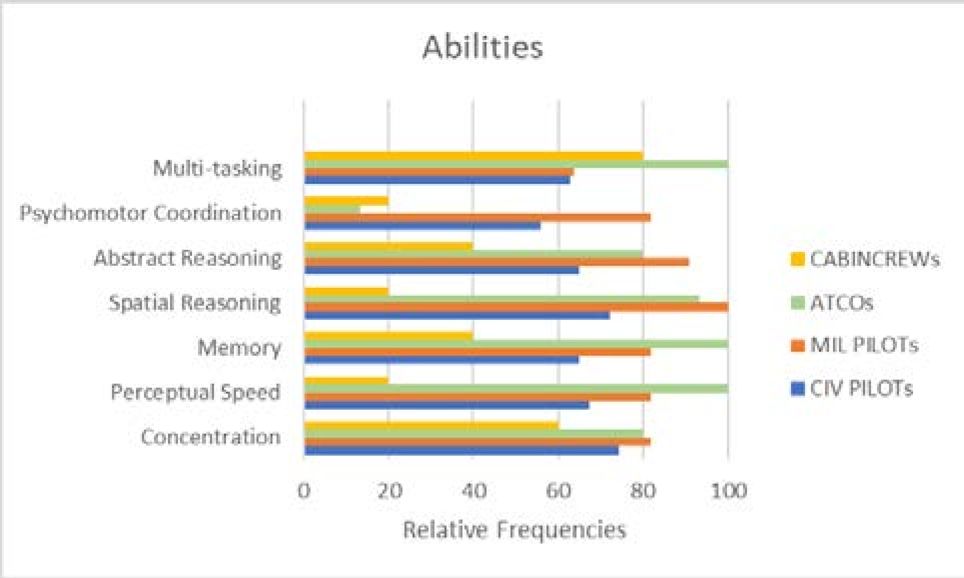

Figure 4: Relative frequencies of ability related selection criteria

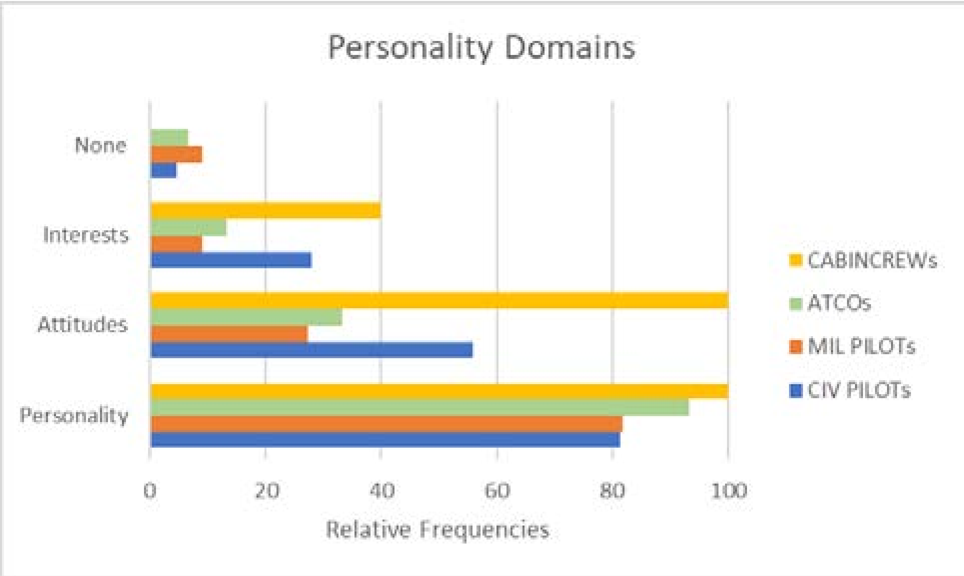

Figure 5: Relative frequencies of personality related selection criteria

According to these figures, the top-rated attributes included by over 80% of the respondents in the respective selection procedures for the different positions are as follows:

- Cabin crews: English knowledge, workload management, communication, teamwork, problem-solving, personality, attitudes
- ATCOs: Perceptual speed, memory, spatial reasoning, multi-tasking, English knowledge, teamwork, communication, motivation, problem solving, decision making, personality
- Military pilots: Concentration, perceptual speed, memory, spatial reasoning, abstract reasoning, psychomotor coordination, English knowledge, motivation, personality
- Civilian pilots: Communication, personality

Assessments of abilities, interpersonal skills, and personality appear to be more common in selection of aviation personnel compared to assessments of knowledge (except for English) or technical skills. This is probably due to the fact that especially ATCOs and military pilots are usually selected at the ab-initio level where technical skills and job-specific knowledge is subject to the subsequent job training.

The rate of agreement for the civilian pilots was the lowest. Only communication and personality were included in most pilot selection procedures. A reason for this disagreement is probably that our survey did not allow a clear distinction between selection systems for ab-initio and experienced pilots. If we lower the threshold from 80% to 70% agreement then the most prominent attributes were concentration, spatial reasoning, English knowledge, leadership, teamwork, communication, problem solving, decision making, and personality.

Table 2 compares the selection methods used to measure the different selection criteria for the four operational groups. The percentages of using computerized classroom tests are higher for selecting ATCOs and military pilots compared to civilian pilots and cabin crews. Interpersonal exercises are used in all groups at a percentage of 49% or higher. These instruments are especially typical for selecting cabin crews. Simulations and work samples are used in less than half of the cases.

	ATCO	Civ Pilots	Mil Pilots	Cabin Crew
Psychometric Tests				
Computerized ability tests	100%	67%	82%	40%
Pen & Paper ability tests	20%	33%	9%	40%
No ability tests	0%	9%	18%	40%
Computerized personality tests	87%	60%	44%	0%
Pen and Paper personality tests	27%	21%	27%	20%
No personality tests	7%	14%	18%	80%
Interpersonal exercises				
Group exercises	67%	49%	64%	100%
Role plays	33%	33%	9%	100%
Work samples and simulations				
Manual control task	7%	23%	36%	0%
Full-flight simulator	0%	21%	18%	0%
Fixed-based simulator	0%	28%	27%	0%
ATC radar simulator	27%	2%	0%	0%
ATC tower simulator	7%	2%	0%	0%
Low fidelity job sample	40%	5%	9%	0%
Interview techniques				
General interview	67%	47%	18%	80%
Psychological interview	80%	70%	82%	80%
Mental health screening	47%	37%	55%	20%

Table 2: Summary of administered selection instruments

Table 3 presents the results of immersive technologies used in selection as well as intentions to use them in future. Less than half of the respondents reported that they are using some kind of immersive techniques. Also, the intentions for their future usage are rather low (≤ 20%).

	ATCO	Civ Pilots	Mil Pilots	Cabin Crew
Immersive techniques applied				
Virtual reality	7%	19%	18%	0%
Augmented reality	0%	9%	9%	0%
Gamification	47%	12%	27%	20%
None	53%	53%	55%	80%
Intentions to use immersive techniques in future				
Virtual reality	20%	12%	0%	20%
Augmented reality	20%	12%	0%	28%
Gamification	13%	9%	9%	20%
None	53%	60%	82%	40%

Table 3: Use of immersive technologies for selection

Length of selection procedures: Respondents were asked how many rounds or stages were in their selection procedure. The specific number of rounds, or stages, in a selection process ranged from 1 to 10 (with an average of 3.4) with 3 rounds being the most frequent answer.

Time taken to assess a successful applicant (i.e. one who is eventually selected for the role) ranged between 2 – 48 hours (average 12.7 hours) with some difference between whether the selection was for civilian pilots (12.2), military pilots (12.3), ATCOs (15) or cabin crew (13.5)

Re-testing: 43.9% of respondents stated that they allow candidate re-testing. Of those, 19.4% allows a re-testing within 1 month, 8.3% within 4 – 6 months, and 72.2% after more than 6 months. For specific organizational groups, 53.3% of answered that they allowed re-testing for ATCO candidates, 41.9% for civilian pilots, 36.4% for military pilots and 0% for cabin crew.

Validation and norms: 80.4% of respondents stated that they had conducted an in-house validation on a selection instrument or battery (93.3% for ATCOs, 67.4% for civilian pilots, 81.8% for military pilots and 40% for cabin crew).

Respondents were asked what norm group they used for assessments, both for personality and ability tests. Options provided were occupation specific in-house norms, occupation specific assessment provider or general population norms.

For personality testing, 59.8% use in-house occupation specific norms, 40.2%

use assessment provider norms, 28% use general population norms. For ability testing 65.4% use in-house occupation specific norms, 40.7% use assessment provider norms, 19.8% use general population norms. A break down by occupational group is provided in table 4.

	ATCO	Civ Pilots	Mil Pilots	Cabin Crew
Personality tests				
In-house occupational specific	66.6%	41.8%	80%	40%
Assessment provider occupation specific	40%	44.2%	0%	0%
General population norms	40%	23.2%	20%	20%
Ability tests				
In-house occupational specific	53.3%	55.8%	70%	80%
Assessment provider occupation specific	66.6%	37.2%	10%	80%
General population norms	20%	16.3%	20%	40%

Table 4: Summary of norm group use by occupational group.

Decision making: 71.1% say they use some form of psychometric cut-off scores when using psychometric tools. 17.2% base their cut-offs on applicants' numbers and available vacancies, 45.6% base it on their own validation research, 33.3% base cut-offs on score distribution and 34.5% use test provider advice for cut-off scores.

	ATCO	Civ Pilots	Mil Pilots	Cabin Crew
Cut-off Scores				
Based on number of applicants	6.7%	18.6%	10%	40%
Based on expectancy of success	60%	34.9%	50%	20%
Based on score distribution	26.7%	25.6%	30%	40%
Based of assessment provider advice	40%	32.6%	20%	20%

Table 5: Summary of use of cut-off scores by occupational group

54.3% respondents state they used composite scores (i.e. an algorithm that yields a total score based on behavioural competency weights or weighted aptitude test scores as some competencies or aptitude tests are more predictive than others) for their selection. Use of composite scores by occupational group was 86.7% for ATCOs, 41.9% for civilian pilots, 50% for military pilots and 60% for cabin crew.

When it comes to the final decision, 24.1% answered that they use algorithmic data integration in their decision making (i.e. test scores and other selection data are statistically combined and integrated for a final selection decision). 75.1% state they use expert data integration (i.e. that test scores and other selection data are combined and integrated based on human judgement for a final selection decision). A breakdown by occupational group is found in table 6.

	ATCO	Civ Pilots	Mil Pilots	Cabin Crew
Integration of selection data				
Algorithmic data integration	20%	22.5%	33.3%	40%
Expert data integration	80%	77.5%	66.7%	60%

Table 6: Integration of selection data by occupational group.

60.8% state that they use compensatory decision making by adding different results in an overall selection/assessment score. 39.2% state they use "hurdling" where single but critical selection/assessment results are used as "knock out" hurdles in the assessment process. Breakdown by occupational group is provided in table 7.

	ATCO	Civ Pilots	Mil Pilots	Cabin Crew
Selection decision making				
Compensatory decision making	80%	72.5%	56.6%	60%
"Hurdling"	20%	27.5%	44.4%	40%

Table 7: Selection decision-making type by occupational group.

Promotions: 45.8% say they use assessments for promotion. Assessments for promotion in 33.3% of ATCO cases and 48.4% of civilian pilot cases. Assessment for promotions was not used for military pilots or cabin crew (0%)

List of tests: EAAP members were asked as part of the survey which assessment instruments or vendors they use as part of their selection process. Note that these are provided in no particular order or categorization and are presented as answered by EAAP members.

The following assessments were provided by respondents: EUROCONTROL's FEAST, Hogrefe, Schuhfried, Vienna Test Systems, Cattell CTI, Testbatair, MMPI-II, NEO-PI-R, CPI, EPQ, PATBA, WOMBAT, PIP, FPR, STAI, AMST System Technique, Par, Naklada SLap, Plutchik PIE Index, 16PF, Rolf Brickenkamp d2 test, Pearson assessments, Psytech, BA, Mindfindr, DLR, Struktúra, Freiburg Personality Inventory, Thinking Schemes Test, Thematic Apperception Test, Raven Progressive Matrix, Advanced Progressive Matrices Short Form, Personal Workplace Values Test, Arctic Shores, Skyrise City, Hogan HPI, Hogan HDS, Saville assessment, Aon cut-e, PCL Risk Type Compass, EQ-i 2.0, SIPISS, Mapi 1/2, Mollymawk psychometrics, COMPASS, GR8 Full Spectrum, Neurolympics (BrainsFirst), Microcog, Eelloo, VCT, Pracownia Testów Psychologicznych PTP, TEA ediciones, Multiplicity, Thomast International, WTS, Bennett Mechanical Comprehension Test, 48QB, GHQ Brief Resilience Scale, Symbiotics, Airways Corporation and in-house developed tests.

14.4. Discussion

According to our findings the most relevant selection criteria in aviation across the four analysed jobs are interactive skills (an average of 78% across the different roles), cognitive abilities (73%), and personality (87%). Less frequently measured attributes are technical skills (49%) and knowledge (52%; except for English with 80%). It could be expected that the importance of technical skills and (job-related) knowledge increases if candidates already had job experience at the time of selection. In our case, especially ATCOs and military pilots are primarily selected at the ab-initio level.

Correspondingly, the most widespread instruments are (computerized) cognitive ability tests, personality tests, and interviews. Interpersonal exercises seem to be on a rise. They had not been identified to a large extent in the IATA and EUROCONTROL surveys. Especially, the selection of cabin crews seems to be primarily based on interpersonal exercises and interviews. Also, the degree of computerization has increased. While EUROCONTROL (2000) reported only 33% usage of computerized ability testing, the figure is now 100% for ATCOs and between 67 and 82%.for pilots.

The acceptance of re-applications has increased for ATCOs from 25% (EUROCONTROL, 2000) to 53% in our study. For pilots our figures are slightly less (between 42% for civilian pilots and 36% for military pilots) than those found by IATA (60%).

A more surprising finding is the relatively rare usage for work samples and simulations. Only between 21% and 40% of the respondents reported using low fidelity job-samples or simulators as selection instruments for ATCOs and pilots. Since meta-studies have repeatedly identified these methods as showing the highest prognostic validities for job performance (Almamari & Traynor, 2019; Martinussen, Jenssen & Joner, 2000), it seems that here is some room for improvement. In this context the application of immersive selection technologies or gamification could enable further developments. At the time of the survey, only a few respondents reported intentions to make more use of such technologies in future. However, this might have been related to onset of the Covid pandemic at the same time.

Indications of scientific rigour can be found in the usage of occupation specific norms (between 40% and 80% for personality and 53% to 80% for abilities) and the realisation of in-house validation studies (80%). Interestingly in this context is that data integration is still based on expert judgement (60% to 80%) and less on algorithmic data integration (20% to 40%). Since meta-studies of mechanical versus clinical decision-making have repeatedly proven the superiority of the mechanical approach (e.g., Kuncel, Klieger, Connelly & Ones, 2013), further efforts to rethink the decision strategies when selecting aviation personnel could be advised.

14.5. Conclusion

The survey results revealed a broad range of how selection in aviation is structured and conducted. While some commonalities exist (e.g. the common usage of ability test, interviews and personality testing), there are also stark differences between selection even within the same occupational group.

More worryingly perhaps, was the widespread use of methods not found in the research literature. This could either be methods that have never been adequately codified or described (e.g. mental health screening or psychological interviews) or methods that have little or no research supporting its use in aviation (e.g. SJT, group exercises or role-play), even though their use in other industries has been examined.

However, this could also be viewed as an opportunity for practitioners in aviation selection. That there are commonly used methods that have no published research data (e.g. SJT for aviation that has little or no published data on efficacy in aviation) opens up the possibility for topical and relevant research and practitioners are urged to use the opportunity to analyze and publish their results.

Moving forward, it is important that EAAP can speak with one voice when discussing or answering questions on aviation selection. The White Paper on Aviation Selection is the first step in that direction, from which quality standards and best practice recommendations can be derived. As psychological selection of aviation personnel becomes mandated in Europe, EAAP may well become the arbiter of what counts as professional or unprofessional when it comes to aviation selection. Disagreements on technical nuances and specific issues in implementation could readily become something that the professional body of aviation psychology would be expected to weigh in on. A survey should this one is an important step to determine the level of professional rigour and where there are opportunities to grow and advance. This type of survey, done regularly, could also become a useful metric to observe how selection changes as time passes and whether quality is being maintained in the aviation system.

14.6. References

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

EUROCONTROL (1997). Information on Available and Emerging Selection Tests and Methods for ab-initio Trainee Controllers (1st ed). Report of the European Air Traffic Control Harmonisation and Integration Programme HUM.ET.ST04.1000-REP-01. EUROCONTROL: Brussels.

EUROCONTROL (2001). Information on Available and Emerging Selection Tests and Methods for ab initio Trainee Controllers (Revised) (2nd ed). Report of the European Air Traffic Management Programme HUM/MSP-002-GUI-01. EUROCONTROL: Brussels.

EUROCONTROL (2000). Update of Information on Selection Tools and Methods for ab initio Trainee Controllers. Report of the European Air Traffic Management Programme HUM/MSP-002-REP-01. EUROCONTROL: Brussels.

International Air Transport Association (IATA) (2019). Pilot aptitude testing - Guidance material and best practices (3rd ed). IATA: Montreal-Geneva.

International Air Transport Association (IATA) (2012). Guidance material and best practices for pilot aptitude testing (2nd ed). IATA: Montreal-Geneva.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. Journal of Applied Psychology, 98(6), 1060–1072

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

ISBN: 9789081525381