A LARGE SCALE VALIDATION STUDY ON AIR TRAFFIC CONTROLLER SELECTION AND TRAINING – DESIGN, CHALLENGES AND RESULTS

Kristin Conzelmann
DLR German Aerospace Center
Hamburg, Germany

Alexander Heintz
DFS Air Navigation Services Academy
Langen, Germany

Hinnerk Eißfeldt
DLR German Aerospace Center
Hamburg, Germany

A validation study with 476 air traffic controller trainees of DFS German Air Navigation Services has been set up, encompassing the complete data from the selection of candidates to the completion of their training. The design includes a detailed coding of interview contents, questionnaire data, and results of the reference sample of 13,716 applicants. Data analysis involves the prediction of training success, training performance, and trainees' personal evaluation of the selection and training process. The success rate of 81 % was satisfactory. Selection measures were adequate to predict pass-fail and performance criteria from institutional training (i.e., theoretical exams). Basic ability measures and a semi-structured interview predicted success in training best. Performance in early training courses was positively related to successive stages and to overall training success. Repetition of exams and switching of training courses/units increased the likelihood of failing. The limits of a correlation-based approach to examine validity are discussed.

Due to the low base rate of suitable applicants, the high safety relevance of the job, and the high costs of ATC training, the selection of ab initio air traffic controller (ATCO) trainees requires a valid and efficient approach to recruitment. The selection procedure for DFS German Air Navigation Services ATC trainees has been developed and conducted in close cooperation with DLR German Aerospace Center. The continuous and intense quality assurance of the process is based on four major pillars, i.e., job requirements analyses (e.g., Bruder, Jörn, & Eißfeldt, 2008; Eißfeldt, & Heintz, 2002), careful selection and regular training of involved psychological and operational staff (Seidel, Pecena, & Eschen-Léguedé, 2009), intense and regular cost benefit analyses (Heintz, 2004), and regular psychometric validation. The selection process involves multiple stages, starting with a paper sift based on published application criteria (e.g., age, education, language training) and a pre-selection based on a biographical questionnaire. The first stage of testing covers a comprehensive set of mental abilities (i.e., memory, concentration, attention, English), followed by a stage of work sample testing (two multiple-task performance tests). The third stage focuses on teamwork abilities (two exercises); the final phase involves an oral English test and a biographical interview including problem-solving tasks. Successful trainees undergo a medical examination according to Eurocontrol class 3 requirements. For each stage and the final selection decision, clearly defined cut-off values and compensation mechanisms are applied.

The validity of the selection program had already been proved in two former large scale validation studies (Damitz, Eißfeldt, Grasshoff, Lorenz, Pecena, & Schwert, 2000; Eißfeldt & Maschke, 1991). These former studies resulted in well-directed adaptations and further developments of the test program (e.g., Pecena, 2003). In international scientific literature, several other validation studies in the field of air traffic controller selection have been reported (for an overview see Broach & Manning, 1998). The current validation study aims primarily at analyzing the complete selection procedure applied during 1997 and 2006 and the reliability and validity of the training process. The major objective of aptitude selection for ATCOs is to increase the probability of success in a safety relevant and costly training. The study at hand was designed to fulfil professional standards of quality assurance and to further enhance the efficiency and validity of the selection and training process.

Method

The sample involves N=476 ATCO trainees (mean age 20.52, S=1.74; 66% male) who were selected between 1997 and 2006. As predictors complete testing data from all selection stages including detailed coding of the

interview content were available. Particularly in the first selection stage, tests scores of the same performance domain can be partly compensated (e.g., a low score in one attention test can be compensated by a higher score of another attention test). Therefore, in addition to single test scores, composite scores of performance domains used for selection decisions were created and used for analysis. In order to apply correction for range restriction, testing data of the corresponding reference sample of N=13,716 test takers was available.

The following criteria were used for analysis: intermediate and final result in terms of pass vs. failure; data of all theoretical and practical exams during institutional training (IT), exam repetition, total duration of training and duration of on-the-job training (OT), and questionnaire data involving self-reports from all phases of selection and training. Compound scores were created of performance assessments and results of written exams. The following composite scores were established: IT overall theoretical and overall practical exam score, composites for each training stage (e.g. initial stages such as "Basic" and "ATC" Course), composite scores of the trainers' ratings in practical exams on detailed performance criteria (e.g. communication, strip handling). See table 1 for a summary of predictors and criteria.

The data set at hand made the data analysis methodologically challenging. Since data was collected over a course of nine years, selection tests were further developed, i.e., item material, count of items, and test duration changed, stanine scores were applied for the major analyses instead of raw scores. The trainees of the validation sample were trained in two different training systems (DATS DFS Air Traffic Controller Training System, DATS 1, N= 430 vs. DATS 2 N=46; DFS, 2010), and for different combinations of licenses (Aerodrome N=91 vs. En route and Approach N=385). Thus, aside from the total sample, several subsamples had to be analyzed taking into account a sufficient sample size. The possibility to cross-check the results across subsamples, was considered as an advantage and yielded additional proof or disproof of the findings. Criterion data was on the whole not normally distributed; trainees were evaluated within a small spectrum of possible grades and got mostly positive ratings. Consequently, non-parametric tests such as Mann-Whitney U-Tests, Spearman correlations and the more complex logistic and ordinal regression analyses were applied to analyze ordinal and categorical data. Logistic regression analysis enabled predicting the dichotomous pass/fail criterion. Structure equation modeling was used to predict the results of later selection stages out of the preceding stage. In additional comprehensive analyses stepwise logistic regression analyses were applied. In the case of data allowing for parametric statistical analysis the corresponding methods were carried out (i.e., Pearson correlations, t-Tests, linear regression, discriminant analysis). Multivariate correction for range restriction was done with the Range J software (Johnson & Ree, 1994) for normally distributed interval-scaled criteria.

Analyses were performed for each selection measure separately. In addition, comprehensive (e.g. multivariate) analyses were calculated in order to include the relationships among the measures. Focus on analysis was the DATS 1 sample since only one trainee of the DATS 2 sample failed. Results were cross-checked with the DATS 1 En route and approach controller sample (=DATS 1, ACC, N=363).

Table 1

Summary of predictor and criterion variables.

Predictor data	**Criterion data**		
			Composite scores
Demographic data (age, sex, education, A-level grade-point-average etc.) Test data • Basic ability tests (concentration, attention, memory, English) • Personality questionnaire • Work sample tests • Team Exercises (group and dyadic) • Oral English exam • (Coded) interview content • Final aptitude level ratings	Institutional Training (IT)	Detailed theory performance assessments in initial training	• Training stages • Overall theory score
		Detailed simulation performance in initial training	• 12 Overall scores on detailed performance ratings
		Detailed results of student license examination	• Overall practical score
		License assignment (Aerodrome vs. en route)	
	Operational Training	OT and total training duration	

Corresponding test data for reference sample (except for interview content)	Pass/fail in intermediate and final check out
Questionnaire on perception of selection and training (only for subsample)	

Main findings

Success rate and training performance. With respect to the success rate, all trainees who failed to complete the training in their initial license assignment (Aerodrome vs. En route) were counted as failure including drop out due to medical reasons or cancellations of training for personal reasons. Overall, 80.7% of the selected trainees of the validation sample successfully validated as ATCOs, corresponding to the regular quality assurance analyses of DFS. The success rate among trainees for aerodrome control towers was higher than for en route and approach controllers (89% vs. 79%). Most failures occurred during operational training (64.8% of all failures). A significantly higher success rate was observed for female trainees, (88.9% compared to 76.4% for male trainees). There were also remarkable differences between the success rates of the DFS units ranging from only 69 % (in one unit) to 100% in some Aerodrome units. While the success rate for 18-19 years old (86.7 % of N=243) and 20 years old trainees (86.0 % of N=157) exceeded the overall success rate, 39% of the trainees with an age of 23 or higher (N=59) failed to validate. Successful trainees had a grade point average (GPA) close to 2 (ranged from 1=very good to 6=insufficient) whereas unsuccessful trainees' GPA approached 3 ($r_{age\text{-}GPA}$=.21; p<.01).

Predictive validity of the selection procedure. Correlations between ability domains of the first selection stage and pass-fail criteria were all positive and significant with respect to concentration, attention and English (r=.08, p<.05 - r=.16, p<.01). All ability domains including the English test correlated consistently with the first training stages (r=.09, p<.05 to r=.29, p<.01). The correlation pattern between the pre-selection tests and the training subjects was positive and mainly significant (r=.09, p<.05 to r=.42, p<.01), too. Concentration and attention correlated significantly with the number of repeated exams (r=.12, p<.05 - r=.13, p<.05, DATS 1, ACC).

Analyses of work sample tests revealed singular significant relationships with subjects of theoretical training (r=.11 - r=.16, p<.01) and with the overall scores on detailed performance ratings (i.e., Traffic planning, Strip Handling, Situational Awareness, Theory, r=.11 - r=.19, p<.05). Team exercise performance revealed some singular significant relationships, for example, with training duration being shorter with a better result in the decision making rating (DATS 1, ACC, r=-.11, p<.05). Neither work sample tests nor team exercises contributed significantly to the prediction of overall training success. This result was surprising since both selection stages proved their validity in the last validation study (Damitz et al., 2000; Höft & Pecena, 2004). A closer look into the data revealed differential predictive validity of work sample test sub scores for male and female applicants, resulting in a differential importance of sub scores for the criteria. These differences are, however, compensated by the overall test score and the significance of the results disappears. Performance in the oral English exam (selection) explained the variance of the English examination result (first training stages) up to 30%. The better the English was judged by the experts, the fewer exams were repeated by the trainees.

In the semi-structured interview the selection board rates an applicant on several dimensions: general motivation, job motivation, cooperation, stress resistance and interactive proficiency. These so-called risk ratings were related to the training criteria. Failure in training was significantly related to a high risk in general motivation (r=-.13, p<.05, particularly in IT), job motivation (r=-.11, p<.05) and cooperation (r=-.13, p<.01, particularly in OT, and student license examinations). A high risk in general motivation increased the probability of exam repetition. The higher the risk was expected to be in interactive proficiency, the longer the training lasted. During the interview, psychologists rate the applicant on specific variables such as parental support, hobbies, efficiency of studies / school. These and additional variables that were hand-coded out of the interview minutes were also related to training success. Correlations were mostly positive and in the expected direction. Particularly, questions about social (e.g., experiences with teachers, relationship to superiors, group membership, self-evaluation) and motivational issues (e.g., efficiency of career, course of studies, hobbies, job motivation) proved to be important predictors of training success in IT and OT (r=.11, p<.05 – r=.23, p<.01).

As a part of comprehensive analyses, stepwise logistic regression analysis was performed with the selection tests on the pass/fail criterion. Concerning the DATS 1 model, 77.4 % of the trainees were assigned correctly to pass and fail (Chi^2=23.78, df=4; Nagelkerke's R^2=.29). Cross checking with the DATS 1, ACC sample revealed an even better classification rate of 82.5 % correct (Chi^2=31.09, df=5; Nagelkerke's R^2=.40). Almost every

selection stage contributed to the model fit with at least one significant test score. Including the aviation specific personality scale that is administered in the context of basic ability testing, the model even improved with respect to the DATS 1, ACC sample (classification correct: 88%; Chi^2=43.94, df=6; Nagelkerke's R^2=.54). The personality scale is only used for interview preparation instead of being applied as a hard criterion in selection. Thus, variance of the personality scale is not yet utilized and has a greater chance to result in significant validation findings. Multiple correlation for the prediction of the overall IT theory score was R=.49 (R=.40 uncorrected, DATS 1, ACC sample). Pre-selection tests predicted the IT theory score best. The overall practical IT score could not be predicted as well as the theoretical score (R=.34; uncorrected: R=.29, DATS 1, ACC sample).

Validity of the training. Training and OT duration were affected by the working position (en-route controllers took longer than aerodrome controllers), change of sector group (e.g., OT duration with change: 21.22 months, S=7.26 compared to 17.54 months without change, S=6.09; Z=-2.66, p<.01), change of training course (e.g., total training duration change excluded: 27.55 months, S=5.53; change included: 32.55, S=5.34; Z=4.37; p<.01). Success rate without changing a training course was 87.9% compared to 64.5% including a switch of training course (N=20; Chi^2 = 13.297, p<.01, w=.17). Results showed a significant tetrachoric correlation between the failure of exams and the pass/fail ratio (r=.41, p<.01). However, 51.7% of the trainees failed in at least one exam during IT, indicating that failing an exam does not necessarily imply total failure. However, without repeating an exam, the success rate was 90% (N=207) compared to 72.2% with resit of exams (Chi^2 = 24.15, p<.01, w=.23). Comparably, fewer repeated practical exams resulted in a higher chance of succeeding in OT and total training (r=-.27, p<.01). Better results (test exams, trainers' evaluations) in IT stages increased the likelihood of OT and total training success (correlations between r=.09, p<.05 and r=.35, p<.01). Within the training stages, there were consistent positive correlations of performance in theoretical test subjects with the pass/fail criterion (with few exceptions, i.e., Aircraft principles of flight, Navigation and English). The trainers' evaluations in the Center course (En route/Approach controller) predicted success in overall training and OT significantly (r=.11- r=.19, p<.05). The better the trainees scored on the twelve DATS criteria (i.e., communication, strip handling) in the practical IT exams, the more likely they finished training successfully.

Perception of selection and training among trainees. The questionnaire reflecting the trainee's perception of various aspects of the selection and training process pointed out relevant insights. The trainees felt they were sufficiently informed throughout the process, they were supported individually to achieve their optimal performance, and both selection and training were appropriate to achieve their objectives in training and in the job. IT and the selection phase received better evaluations compared to the OT phase. Women stated that they received more feedback in IT (M=3.44, S=.66) and OT (M=3.62, S=.56) compared to men (M=3.07, S=.81 and M=3.39, S=.66, all differences p<.05). Male trainees, however, found the transition to simulation training easier than women (M=3.41, S=.71 compared to M=3.09, S=.88, p<.05). Ratings differed also according to the unit. Failed trainees evaluated OT worse than trainees who validated.

Discussion

The overall training success rate was sufficient and increased compared to former evaluations. However, remarkable differences concerning the gender of the trainees have to be explored further. Evidence for factors beyond the ability level is under examination in order to identify ways to enhance the success rate of male trainees. The educational level and age are significantly related to training success, which confirm the relatively strict application of criteria for ATCO applicants at DFS, i.e., accepting only applicants with A-level exam who should not be older than 24 years. Operational units can be objectively informed on the impact of failed exams during initial training and their limited relationship to failure in the operational training. This helps to prevent a "Pygmalion" effect negatively impacting the attitude towards trainees in the OT phase (Rosenthal & Jacobson, 1992).

Concerning the selection process, as a main consequence of the study and to further increase the success rate, one should focus on tests which are not already used as hard or explicit criteria in the selection procedure in order to make better use of additional sources of variance. For example, some personality scales and specific categories of questions within the semi-structured interview yielded additional gain in predicting training success. Despite comparatively low correlation effects of test adaptations based on former studies, the increase of the training success rate confirms the success of these adaptations. Correlation-based validation approaches in selection processes with a very low variance in predictor and criterion data are limited. The selection cut-off values (selection rate of 5-6%) and the limits for high safety related training performance are comparably strict; usually "false positive" selection decisions are avoided. Without considering the increase of the training success rate, there would be a risk of underestimating selection tests and training exams that did not prove to be valid in terms of correlation-based analyses within the selected group but probably affect the success rate in a positive

way. Moreover, there is a strong risk of overestimating tests which reveal high correlations with training performance irrespective of the training success rate of trainees selected based on these tests.

References

Broach, D. & Manning, C. (1998). Issues in the selection of air traffic controllers (pp. 237-271). In M. W. Smolensky & E. S. Stein (Eds.), *Human factors in air traffic control*. San Diego, CA: Academic Press.

Bruder, C., Jörn, L. & Eißfeldt, H. (2008). Aviator 2030: *When pilots and air traffic controllers discuss their future*. In A. Droog & T. D'Oliveira (Eds.) Proceedings of the 28th EAAP Conference, Valencia, Spain (Vol 2 pp.354-384). Valencia: European Association for Aviation Psychology.

Damitz, M., Eißfeldt, H., Grasshoff, D., Lorenz, B., Pecena, Y. & Schwert, T. (2000). *Validierung des DLR-Auswahlverfahrens für Nachwuchsfluglotsen der DFS Deutsche Flugsicherung GmbH: Ergebnisse des Projektes Qualitätssicherung*. DLR FB 2000-45. Hamburg: DLR.

DFS German Air Navigation Services (2010). *DFS Initial Training for ATCO Syllabus and Training Event Plans. Version 1.1*. Langen: DFS.

Eißfeldt, H. & Heintz, A. (2002). *Ability requirements for DFS controllers - Current and future.* In H. Eißfeldt, M. C. Heil & D. Broach (Eds.), Staffing the ATM system - the selection of air traffic controllers (pp. 13-24). Aldershot: Ashgate.

Eißfeldt, H. & Maschke, P. (1991). *Bewährungskontrolle eines psychologischen Auswahlverfahrens für den Flugverkehrskontrolldienst anhand von Kriterien der Berufsausbildung*. DLR-FB 91-11. Hamburg: DLR.

Heintz, A. (2004). *Cost-benefit analysis in the selection of Air Traffic Controllers.* In Proceedings of the 26th EAAP Conference - Aviation Psychology: Costs and Benefits. Sesimbra: European Association for Aviation Psychology.

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.

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

Pecena, Y. (2003). *An Assessment-Center Approach to ATCO Selection – An Evaluation Study*. In Proceedings of the second EUROCONTROL Selection Seminar (pp. 201-208). Luxembourg, 16.-18.Nov.2003.

Rosenthal, Robert & Jacobson, Lenore (1992). *Pygmalion in the classroom.* Expanded edition. New York: Irvington.

Seidel, K., Pecena, Y., & Eschen-Léguedé, S. (2009). *Computer-based assessor training - a possibility to solve the dilemma of economical efficiency and quality?* In Proceedings of the 14th European Congress of Work and Organizational Psychology, Santiago de Compostela, Spain, 13.-16. May 2009.