A Psychometrician's Dream: Test Validities in an Unrestricted Sample of Future Swiss Air Force Pilots

Hans-Juergen Hoermann[1)], Philip Noser[2)] & Dirk Stelling[1)]

[1]Aerospace Psychology, German Aerospace Center (DLR), Hamburg, Germany

[2]Institute of Aviation Medicine (FAI), Swiss Air Force, Duebendorf, Switzerland

Abstract.

Introduction. In the context of personnel selection the predictive validities of the selection tests are usually based on the restricted sample of recommended applicants. The resulting validity coefficients can be corrected by the restriction-of-range formula, which is based on a number of assumptions especially in multivariate settings. **Research question.** Which tests show the highest predictive validity for flying performance in an unrestricted sample of student pilots? **Method.** N=135 Swiss Air Force applicants were examined with the DLR selection battery. Criterion information about flying performance was collected after ten flying lessons. **Results.** A number of substantial predictor-criterion correlations were found. **Discussion.** The highest predictive validities were found for psychomotor tests under multitasking conditions, spatial orientation, and mechanical comprehension. **Conclusion.** The findings confirm results from earlier studies showing that complex psychomotor tests are the best predictors for the first stages of training to become a successful pilot.

Keywords: pilot selection, psychomotor performance, multi-tasking, predictive validities, restriction of range

Introduction

Predictive validity is one of the most important quality standards for psychometric selection tests. However, a simple predictor-criterion correlation often underestimates the strength of this relationship, because normally only data from parts of the original sample are available. For applicants who did not pass the selection tests, generally no further feedback information exists. This situation causes a restriction of range in the distribution of the selection test scores when applicants with low performance scores are washed out during this process. Therefore, such a restricted sample and any calculated test statistics would not be representative for the entire population of applicants. Especially, those tests with a high weight and therefore narrow range of acceptable scores will show the strongest effects due to restricted variances, which can lead to a misjudgement of the predictor-criterion relationship in the population.

As a countermeasure, correction formulas had been published to compensate the respective correlations coefficients for the restricted variances (e.g. Lawley, 1943; Gulliksen, 1987; Ree, Carretta, Earles, and Albert, 1994). However, these correction formulas have a number of limitations. First, corrected correlations cannot be tested for significance. Therefore, it depends on the user, which size of a corrected correlation is regarded as noteworthy. Second, in a multivariate, multistage setting the test variances of a new test can be restricted even if the test scores were not used for selection. This is caused by the usual intercorrelation between different predictor tests or by a preselection prior to the actual selection tests. For example, the predictive validity of a university entrance test in Germany was underestimated, because as a condition to take part in this test applicants had to demonstrate a certain level of

school achievements. Since the admission test was correlated with the school grades, the possible test variance was already restricted even before the admission test had been administered (Bartussek, Raatz, Stapf, and Schneider, 1986). The Lawley (1943) approach accommodates for multivariate settings, but only for the bivariate correlation coefficients and not for a multiple regression approach. Third, the correction formulas require an estimate of the predictor distribution in the unrestricted population. The estimation of the predictor variances in the unrestricted population is difficult to obtain especially when the time-span for administering the predictor tests is large and varies for the different predictor tests.

Predictive validity studies in aviation based on unrestricted samples are extremely rare in the published literature. A classic example was reported by Thorndike (1949) with $N = 1036$ Air Force pilot applicants who took a number of selection tests, but afterwards all applicants were admitted to pilot training regardless of their test performance. As Thorndike described, the predictive validity for a complex psychomotor coordination test with pilot training success in the unrestricted sample was $r = .40$. The same relationship changed to $r = -.04$ when through the application of selection criteria the sample had been restricted to $N = 136$ (selection ratio = 13.1%). For another pilot specific predictor test of mechanical comprehension the correlation attenuated from $r = .46$ in the unselected sample to $r = .03$ in the sample of potentially qualified applicants. This shows how meaningless correlation coefficients can become if they are based on a restricted sample without application of correction formulas.

In this research paper the predictive validities of a battery of typical pilot selection tests will be examined in an unrestricted sample of candidates. The only limitations were Swiss citizenship and the decision to apply as a pilot for the Air Force based on self-selection so to speak. Evaluations of flying instructors after a two-week flight training course were adopted as criteria. Additionally, the attenuation of predictive correlation coefficients will be compared across two further sub-samples restricted by application of a lenient or of a strict selection procedure.

Method

Sample: Test data and criterion information was available from $N = 135$ pilot applicants for the Swiss Air Force. The average age was 18 years (range 17 to 20 years). 89.6% candidates were male. Applicants had to be Swiss citizens of good health between 17 and 22 years old when the selection tests were administered. No specific educational degree was required (see SPHAIR website, 2018). All 135 candidates participated unfiltered during the year 2010 in the initial flying appraisal within the SPHAIR program (see Noser 2011 and Noser and Laege, 2012) although $N = 30$ (22.2%) did not pass the prior selection tests according to the lenient selection procedure. The suspension of selection at this stage was justified because of the introduction of a new psychomotor and multitasking test (PMA, Noser, 2011).

Predictor tests: The selection tests were administered on computers during one-day test sessions. The test battery included the following tests:

- ENS: Written English comprehension test
- TVT: Mechanical comprehension test
- RAG: Mathematical reasoning test
- KRN: Mental arithmetic test
- MST: Memory search task
- OWT: Optical perceptual speed test
- SKT: Mental concentration test

- ROT: Mental rotation test
- MIC: Monitoring and instrument coordination test
    - MIC-1: Psychomotor tracking sub-score
    - MIC-2: Total score (three tracking tasks plus auditory monitoring)
- PMA: Psychomotor and multitasking test
    - PMA-1: Memorization sub-score
    - PMA-2: Information ordering sub-score
    - PMA-3: Selective attention sub-score
    - PMA-4: Psychomotor tracking sub-score
    - PMA-5: Total score (three tracking tasks plus three cognitive tests)

The first four tests (ENS, TVT, RAG, KRN) can be regarded as general scholastic aptitude tests. ENS and TVT are multiple choice tests. For RAG and KRN the calculation results had to be entered via a numerical keypad. MST, OWT, SKT, and ROT are tests of cognitive abilities more related to specific task requirements of pilots. The MIC (Hoermann, 2016) and the new PMA (Noser, 2011) are complex tests of psychomotor coordination, scanning abilities and time sharing capacity. These two tests offer different sub-scores plus an aggregated total score (MIC-2 and PMA-5).

Besides the ability tests the Temperament Structure Scales TSS (Maschke, 1987), a personality questionnaire was administered with the subscales:

- ACH: Achievement motivation
- RIG: Rigidity
- VIT: Vitality
- EXT: Extraversion
- DOM: Dominance
- EMP: Empathy
- AGG: Aggressiveness
- EIN: Emotional instability
- OPN: Openness

Criteria: The flying appraisal is part of the regular selection process for Swiss pilots. Normally, only those candidates progress into this selection stage if their prior test scores were positive. For the purpose of this study all candidates were approved for the flying lessons regardless of the test performance. The flying appraisal was based on a two-week standardized flight training consisting of theoretical instruction and eleven flying lessons conducted in a flight training centre in Switzerland. Nine performance aspects (situation awareness, decision making, visualisation, motor control, handling, adherence to rules, memory, stress resistance, stamina) and seven personality aspects (self-assessment, behaviour, communication, agreeableness, initiative, reliability, frustration tolerance) were graded on four-point scales by familiarized flight instructors. The test instructors were kept uninformed about the prior test results. Two total evaluation scores were available for flying performance and for personality.

Procedure: The full sample of 135 candidates was restricted by application of a lenient and by a strict selection process. With the lenient selection process the sample was restricted to N = 105 (Sample 1), which corresponds to a selection ratio of 77.8%. With the strict selection process the sample was restricted to N = 52 (Sample 2), which corresponds to a selection ratio of 38.5%.

Results

Pearson product-moment correlations were calculated to predict the final performance evaluation of the flight instructor with the individual selection tests. First, these correlations were calculated in the full sample and afterwards in the two restricted samples. All coefficients were neither corrected for criterion unreliability nor for the range restriction effects. The results are shown in Table 1 below.

Table 1. Predictive validities for the instructor evaluation of overall flying performance in the unrestricted sample and in two restricted sub-samples

Predictor	Full Sample (N = 135)	Restricted Sample 1 (N = 105)	Restricted Sample 2 (N = 52)
English Language	.20*	.16	.11
Mechanical Comprehension	.33**	.29*	.44**
Mathematical Reasoning	.23**	.20*	.38**
Mental Arithmetic	.23**	.14	.14
Memory Search	-.04	-.16	-.04
Optical Perceptual Speed	.20*	.15	.03
Mental Concentration	.14	.02	-.24
Mental Rotation	.34**	.23*	.21
MIC-1: Psychomotor Tracking	.24**	.15	.39**
MIC-2: Multi-tasking	.52**	.51**	.62**
PMA-1: Memorization	.15	.05	.01
PMA-2: Information Ordering	.21*	.13	.12
PMA-3: Selective Attention	.09	.08	.01
PMA-4: Tracking	.26**	.17	.35*
PMA-5: Multi-tasking	.33**	.22*	.27

Notes. ** $p < .01$; * $p < .05$

As the results of this correlation analysis show, most of the predictors did have significant predictive validities with the criterion measure of flying performance. The highest validities of $r \geq .30$ were found for the Mechanical Comprehension Test, Mental Rotation, and Multi-tasking. Of really substantial size is the correlation for the multi-tasking score of the MIC. Memory Search, Mental Concentration, and Selective Attention did not contribute any significant variance portion of this criterion.

The two columns in the right half of Table 1 show the results for the restricted samples. Even though the selection ratio was quite high when applying the more lenient selection process, all correlation coefficients were lower and most of them became insignificant in the restricted sub-sample 1. However, as can be seen for sub-sample 2, the further restriction to about 38% of the original sample did not cause a further shrinkage of the predictive validities. In contrary, for the predictor tests with the highest validities the coefficients remained rather stable or even slightly increased. The inspection of the standard deviations confirmed the range restriction. For example, the standard deviations in the three samples for Mechanical Comprehension were $s = 1.95$ (full sample), $s = 1.94$ (sample 1), $s = 1.35$ (sample 2). The standard deviations for MIC-2 were $s = 1.98$ (full sample), $s = 1.77$ (sample 1), $s = 1.54$ (sample 2). Figure 1 shows the scatterplots for the Mechanical Comprehension test in the full

sample and in the most restricted sample 2. Actually, the residuals seem to decrease with the size of the Mechanical Comprehension stanine, which means that the predictor-criterion relationship may be lacking homoscedasticity in the full sample.

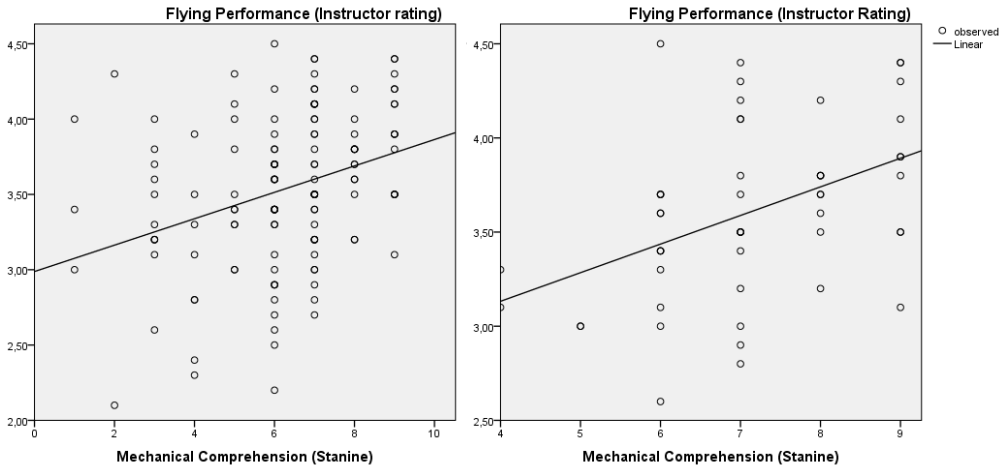

Figure 1. Scatterplots for the Mechanical Comprehension test in the full sample (left) and in sample 2 (right)

With multiple regression analysis the predictive power of all fifteen selection tests for the overall evaluation of flying performance by the flight instructors were examined. For the unrestricted sample the multivariate correlation was regression $R = .60$** (corrected *R-Square* = .27). However, the only significant predictors were Mechanical Comprehension and MIC-2 Multi-tasking. With a lenient selection process we found in sample 1 $R = .60$** (corrected *R-Square* = .25) and with a strict selection process we found in sample 2 $R = .80$** (corrected *R-Square* = .47). Only MIC-2 Multi-tasking remained as a significant predictor in the latter two regression equations.

None of the TSS dimensions correlated significantly in the full sample with the overall personality evaluations of the flight instructors. Also the correlations with the seven personality sub-aspects were mostly insignificant except for the relationship between achievement motivation (TSS-ACH) and initiative ($r = .24$*), for dominance (TSS-DOM) with communication ($r = .32$**), and for aggressiveness (TSS-AGG) with communication ($r = .24$*). Instead of correlations with the TSS dimensions the personality evaluations of the flight instructors related to some performance scores of the selection tests.

Discussion/Conclusions

The analysis of predictive validities of pilot selection tests in an unrestricted sample of Swiss Air Force pilots were examined in this study. The objective was to identify the most relevant predictors for flight instructors' evaluations of flying performance subsequent to initial flight training. In the unbiased full sample of candidates almost all performance tests except those for memory and selective attention demonstrated significant bivariate correlations with the criterion. Clearly the best predictors were the complex psychomotor tests MIC and PMA especially with the sub-scores for multi-tasking. This was confirmed by multiple regression analyses, which resulted in significantly high multiple correlations of $R = .60$ and higher.

The effects of increasing range restriction by application of a lenient and a strict selection process on the size of the correlation coefficients were less than expected. Though the standard deviations reflected the restriction of range by application of the selection rules, only some of the correlation coefficients decreased in size in the reduced samples. Correlations for

the best predictors remained significant even after 22.2% to 61.5% of the candidates were selected out. As illustrated in Figure 1 a lack of homoscedasticity for the predictor-criterion relationship could be a reason that range restriction did not affect all correlations equally.

While the prediction of flying performance with the selection tests was apparently demonstrated, personality evaluations of the flying instructors could not be predicted with the personality inventory TSS. In this early stage of flight training the instructor ratings were influenced more by the pure flying performance and obviously less by the social behaviour of the candidates as shown by the correlations of the personality evaluations with performance tests. Helmreich, Sawin, and Carsrud (1986) called this the "honeymoon effect", which means that the influence of personality on a pilot's career would become more salient in later career stages only.

In this study we have examined only an intermediate criterion of initial flying performance. The predictive validities for the selection tests of performance especially the complex psychomotor tests were clearly demonstrated. However, with this data set we could not show how accurate pilot performance in later career stages can be predicted. This should be investigated in follow-up studies with the successful student pilots of this sample.

References

Bartussek, D., Raatz, U., Stapf, K.H., & Schneider, B. (1986). *Die Evaluation des "Tests für medizinische Studiengänge"*. Zweiter Zwischenbericht. Bonn.

Gulliksen, H.(1987). *Theory of mental tests*. New York: Lawrence Erlbaum Associates.

Helmreich, R.L., Sawin, L.L., Carsrud, A. (1986). The honeymoon effect in job performance. Temporal increases in the predictive power of achievement motivation. *Journal of Applied Psychology*, *71*, 185-188.

Hoermann, H.-J. (2016). *MIC: Monitoring & Instrument Coordination – Documentation*. Unpublished report, DLR, Hamburg.

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.

Noser, P. & Laege, D. (2012). Psychomotor function and multitasking abilities: Development of a new test system for pilot aptitude prediction in Swiss Air Force. *Poster presented at the 30th EAAP Conference 24-28 September*. Villasimius, Sardinia/Italy.

Noser, P. (2011). *Eignungsdiagnose von Militärpiloten: diagnostischer Rahmen und Messung von Multitasking-Fähigkeiten*. Dissertation, Universität Zürich, Psychologisches Institut, Abt. Allg. Psychologie (Kognition).

Ree, M.J., Carretta, T.R., Earles, J.A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology*, *79*, 298-301.

SPHAIR (2018). *SPHAIR Prozess Jet-Pilot*. Retrieved from https://www.sphair.ch/sphair/der-sphair-prozess-pilot-jet .

Thorndike, R.L. (1949). *Personnel selection; test and measurement techniques*. New York: Wiley.

Contact Information:

Dr. Hans-Juergen Hoermann
German Aerospace Center (DLR), 22335 Hamburg, Sportallee 54a
Hans.hoermann@dlr.de