

VALIDATION OF THE DUTCH AIRFORCE TEST BATTERY USING ARTIFICIAL NEURAL NETWORKS

Markus Sommer, Joachim Häusler, Koning, A. J. and Martin Arendasy

*Dr. G. Schuhfried GmbH
Hyrtlstr. 45
2340 Mödling
Austria*

sommer@schuhfried.at

THEORETICAL INTRODUCTION

The main selection criteria for individual tests and test batteries used to select military pilot applicants are the construct and criterion validity, the overall cost of testing and the time requirements. Naturally, the derivation of decisions from a test battery requires a sufficiently high correlation between the tests and the criterion variable. However, recent metaanalysis (cf. Burke, Hobson & Linsky, 1997; Hunter & Burke, 1994; Martinussen, 1996) indicates that the correlation coefficients between a single test and the criterion measure do not exceed an absolute value of .30. There are a variety of causes for this, ranging from a lower reliability of the criterion or predictor variables (Lienert & Raatz, 1998; Goeters, 1998), an attenuation of the variance in the predictor variables due to selection (Lienert & Raatz, 1998; Goeters, 1998) to the lack of symmetry between the generality of the predictor variables and the generality of the criterion variable. With regard to the later cause Wittmann and Süß (1997), Ajzen (1987) and Ree and Carretta (1996) pointed out that for more general and global criteria such as successful performance in a flight-simulator or an educational program, aggregate measures such as general ability ("g") are better suited for prediction than more specific predictors. Thus one way to handle this problem is to combine the available information about an applicant to generate a prediction about his success. In general, one can resort to various methods of statistical judgment formation in order to do so. But classical methods of statistical judgment formation such as discriminant analysis or regression analysis are vulnerable to violations of their statistical assumptions and often lack stability in cross-validation in practical applications (cf. Bortz, 1999; Brown & Wickers, 2000). A promising alternative is the use of artificial neural networks. This statistical method has few requirements with respect to data characteristics and has proven to be a robust procedure for pattern recognition tasks (Bishop, 1995; Kinnebrock, 1992; Mielke, 2001; Rojas, 2000; Warner & Misra, 1996). In a previous study Griffin (1998) evaluated artificial neural networks with regard to their ability to predict naval aviator flight grades in their primary phase of flight training using a test battery which primarily consisted of psychomotor tests. Griffin's results indicated that artificial neural networks resulted in a higher validity coefficient compared to the multiple linear regression analysis. However the difference did not reach statistical

significance. In line with the current literature on neural networks (Bishop, 1995), the author attributed this result to the lack of non-linear relations between the chosen predictor variables and the criterion variable. Based on this result the aim of the present study is to compare linear discriminant analysis and a neural network with respect to classification rate and generalizability.

METHOD

The first stage of the selection procedure for pilot applicants involved the use of psychological tests. The tests for the psychological dimensions mentioned in the JAR-FCL3 were selected based on their theoretical foundation and construct validity. The test battery consisted of the following subtests taken from the Intelligence Structure Battery S2 (INSBAT: Arendasy, Hornke, Sommer, Häusler, Wagner-Menghin, Gittler, Bogner & Wenzl, 2005): Numerical-inductive reasoning (NID), Figural-inductive reasoning (FID), Arithmetical competence (AK), Computational estimation (ASF), Numerical flexibility (NF), Inspection time (BZ) and Decision quality and speed (EF). The first two subtests measure different aspects of the second stratum factor fluid intelligence (G_f), while the third, fourth and fifth subtests assess individual differences in different facets of the stratum two factor quantitative reasoning (G_q). The subtests Inspection time (BZ) and Decision quality and speed (EF) measure the aspects of decision speed (G_{ds}). Furthermore, the Adaptive Three-Dimensional Cube Test S2 (A3DW: Gittler, 1998) was used to measure spatial rotation, while Cognitron S4 (COG: Wagner & Karner, 2003) was used for measuring selective attention. In order to measure perceptual speed the Tachistoscopic Traffic Perception Test S1 (TAVT: Biehl, 2002) was also administered.

In the case of the INSBAT subtests as well as A3DW and TAVT the person parameters in accordance with the Rasch model (Rasch, 1980) obtained by the respondents in the respective test were included in the analysis. In the case of COG the main variables “sum of correct reactions” was used as the predictor variable.

A second selection phase involved global assessments of the subjects’ performance in a standardized flight simulator. The global assessment of the trainees’ performance in the simulator served as a criterion variable. On the basis of the global assessment of their performance in the flight simulator the respondents were subdivided into a group of successful and not successful military pilot applicants.

Sample

The sample encompasses 150 pilot applicants for the Dutch Airforce. The complete data of 99 pilot applicants are provided. The remaining pilot applicants did not complete the entire test battery and were thus excluded in the multivariate analysis. The remaining sample consists of 98 (99.00%) male pilot applicants and one (1.00%) female pilot applicant. All the candidates are between 16 and 25 years of age, with a mean age of 18.84 years and a standard deviation of 2.04 years. A total of 61 (61.6%) military pilot applicants received a negative global evaluation of their performance in the standardized flight simulator.

RESULTS

Results obtained with non-linear methods:

The artificial neural network was calculated using the program NN Predict (Häusler, 2004). The type of network used consisted of a multi-layer perceptron with one functional intermediate layer and full feed-forward connection. As a transformation function the activation function Softmax was used, which represents in essence a "multiple logistical" function, the result of which can be interpreted as a posteriori probability. According to Bridle (1990), this activation function is particularly suitable for use with categorical criterion variables. QuickProp (Fahlmann, 1988) was used as the learning algorithm. The number of iterations was 5000. Following a suggestion of Häusler and Sommer (2006), the number of predictor variables and intermediate layer elements was determined by comparing different network architectures with varying numbers of intermediate layer elements on the basis of their adjusted validity coefficient (adj. R²) and economy (BIC). The results are provided in table 1.

Table1: Predictor variables, number of hidden layer units, BIC and adj. R² for different artificial neural network architectures.

Predictor variables	hidden layer elements	BIC	adj R ²
NID, AK, BZ, EF, FID	3	348.5	.462
NID, ASF, AK, EF, NF, A3DW	2	351.7	.256
NID, AK, BZ, EF, FID, NF, TAVT	3	345.22	.599

As can be seen in table 1 the total optimization resulted in an optimum number of three hidden layer elements and a total of seven predictor variables taken from the subtests Numerical-inductive reasoning (NID), Arithmetical competence (AK), Figural-inductive Reasoning (FID), Numerical flexibility (NF), Inspection time (BZ), Decision quality and speed (EF) and the main variable overview of the Tachistoscopic Traffic Perception Test (TAVT-UEB).

Using the empirically derived number of hidden layer elements and the empirically derived predictor variables, the predictive validity of this optimized test battery was investigated. Table 2 summarizes the validity coefficients and classification rates for the simple classification and for the jackknife validation.

Table 2: Validity (R), adjusted explained variance of the criterion (adj. R²), classification rate (CR), sensitivity (1-β) and specificity (1-α) of the prediction after simple training of the artificial neural network and according to the jackknife method for the optimized test battery, based on the total sample. The validity was calculated as the correlation between true value and predicted value.

Simple prediction					Jackknife prediction				
R	adj. R ²	CR	1-β	1-α	R	adj. R ²	KR	1-β	1-α
.84	.61	92.9	84.2	98.4	.83	.59	92.9	84.2	98.4

As can be seen from Table 2, the results regarding the predictive validity of the optimized test battery are confirmed by the jackknife method. The validity coefficients and classification rates, both for the simple prediction and in the jackknife validation, are high, with a reasonably balanced relationship between sensitivity and specificity. The correlation between the classification probabilities of the simple prediction and the jackknife validation is R=.99. The results therefore reveal a substantial correspondence between the simple prediction and the jackknife validation, which indicates that the stability of the results is high. In order to ensure the stability of the results we further validated the model by means of an internal bootstrap. This involved setting up the model using the complete data set and then testing it on 1000 bootstrap samples. The validity coefficient and the classification rate were calculated for each bootstrap sample. For the validity coefficient a confidence interval of [.74; .94] was obtained, while the confidence interval for the classification rate was [88.2 %; 97.7 %]. In summary it can be said that the results of both the bootstrap and the jackknife validations indicate that the network architecture used in this study provides a stable result.

The next step involved calculating the incremental validity and the relative relevance of the individual test variables of the optimized test battery. The results are presented in table 3.

Table 3: Incremental validity and relative relevance of the main variables from the optimized test battery

Predictors	Incremental validity	Relative relevance
Numerical-inductive reasoning (NID)	.157	13.4 %
Arithmetical competence (AK)	.205	16.9 %
Inspection time (BZ)	.195	16.2 %
Decision quality and speed (EF)	.167	14.1 %
Figural-inductive reasoning (FID)	.324	24.6 %
Numerical flexibility (NF)	.105	4.6 %
Overview (TAVT)	.117	10.2 %

As can be seen in Table 3 the two subtests Numerical-inductive reasoning (NID) and Figural-inductive Reasoning (FID) contribute the most to the predictive validity of the optimized test battery. This result argues for the importance of fluid intelligence (G_f) in predicting the success of pilot applicants. Arithmetical competence (AK), Inspection time (BZ), Decision quality and speed (EF) and the main variable overview from the Tachistoscopic Traffic Perception Test (TAVT-UEB) also contribute substantially to the predictive validity indicating

the importance of quantitative reasoning (G_q) and mental speed (G_s) in selecting pilot applicants. Numerical flexibility (NF) contributes less than the other predictors but nevertheless proved to contribute significantly to the predictive validity of the optimized test battery. The result is in accordance with the relevance attributed to quantitative reasoning in the JAR-FCL. Contrary to our prior assumptions the Adaptive Three-Dimensional Cube Test (A3DW) and Cognitrone (COG) did not significantly contribute to the predictive validity of the test battery above and beyond the predictor variables included in the optimized test battery.

Where practical applicability is concerned, the level of certainty of the classifications on the individual subject level is also of importance. This can be investigated using the distribution of the classification probabilities calculated with the aid of the artificial neural network. Figure 1 shows the distribution of the estimated person-specific probability of being classified as passing the flight simulator test. The x-axis shows the success probability. For the sake of clarity the person-related probabilities were summarized in ten groups. The y-axis represents the proportion of individuals who actually received a positive (white bar) or negative (black bar) evaluation.

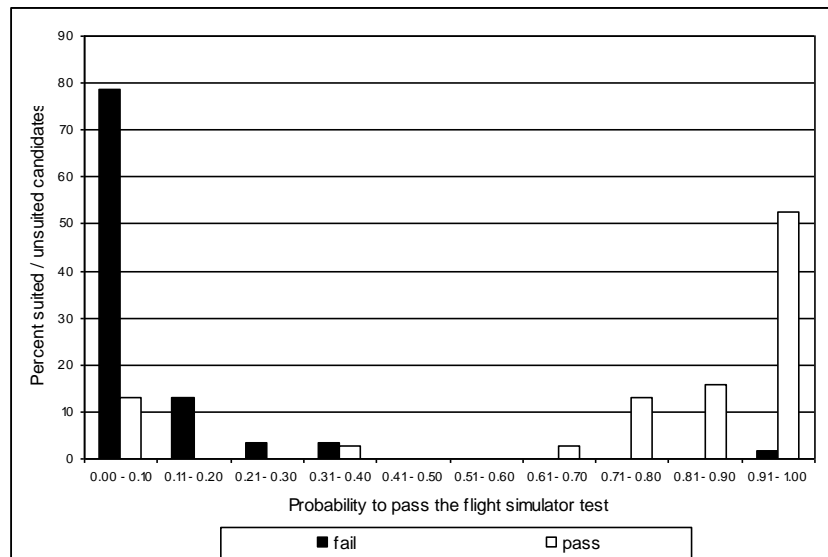


Figure 1: Classification of a trained artificial neural network according to the jackknife method on the basis of the result in the optimized test battery. The x-axis shows the probability of passing the flight simulator test; the probabilities are divided into ten groups. The bars indicate the percentage in each group of subjects who actually failed (black bar) or passed (white bar) the flight simulator test.

In summary it can therefore be said that even in the case of the assessment of individual cases the optimized test battery achieves a good discrimination between successful and less successful pilot applicants.

Results obtained with linear methods

The calculation of the discriminant analysis was carried out with SPSS 14. The results indicate that the discriminant analysis is unable to separate successful and less successful pilot applicants based on their test scores (Wilks-Lambda=.927, df=7, p=.418; Box-M: F=1.421, p=.069). Using the empirically derived predictor variables by means of the artificial neural network, the predictive validity of the optimized test battery was investigated. Table 4 summarizes the validity coefficients and classification rates for the simple classification and for the jackknife validation.

Table 4: Validity (R), adjusted explained variance of the criterion (adj. R²), classification rate (CR), sensitivity (1-β) and specificity (1-α) of the prediction for the entire data set and according to the jackknife validation of the discriminant analysis. The validity was calculated as the correlation between true value and predicted value.

Simple prediction					Jackknife prediction				
R	adj. R ²	CR	1-β	1-α	R	adj. R ²	KR	1-β	1-α
.264	.070	58.6	57.4	60.5	.205	.042	48.5	47.5	50.0

As can be seen in Table 4, the results regarding the predictive validity of the optimized test battery using a linear method are considerably lower than the ones obtained with artificial neural networks. Furthermore, the correlation between classification probabilities of the simple prediction and the jackknife validation amounts to R=.61. It can thus be concluded that the results obtained with the discriminant analysis are less stable than the results obtained with artificial neural networks. The results were further validated means of an internal bootstrap. This involved the estimation of the model parameters of the discriminant analysis in the whole data set and then testing it on 1000 bootstrap samples. The validity coefficient and the classification rate were calculated for each bootstrap sample. For the validity coefficient a confidence interval of [.12; .52] was obtained, while the confidence interval for the classification rate was [53.0 %; 73.4 %]. Compared to the results obtained with artificial neural networks the confidence intervals for both the classification rate and the validity coefficient are quite large. Taken together the results indicate that the solution obtained by means of a linear discriminant analysis proved to be far less stable than the results obtained by means of an artificial neural network.

Figure 2 shows the distribution of the estimated person-specific probability of being classified as passing the flight simulator test according to the discriminant analysis.

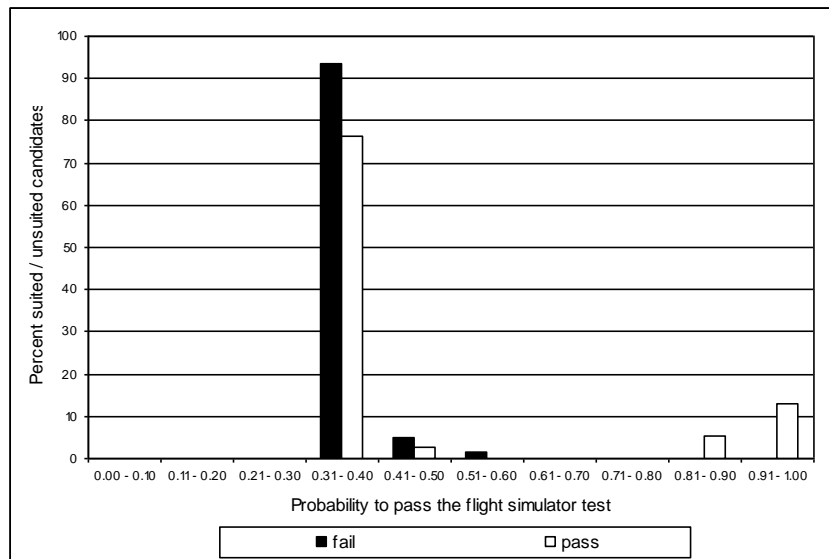


Figure 2: Classification based on the jackknife validation of the results obtained with a discriminant analysis on the basis of the result in the optimized test battery. The x-axis shows the probability of passing the flight simulator test; the probabilities are divided into ten groups. The bars indicate the percentage in each group of subjects who actually failed (black bar) or passed (white bar) the flight simulator test.

As can be seen in Figure 2 the individual classification probabilities are close to a chance rate of .50. This reflects the inability to separate able and less able pilot applicants based on test scores when using linear classification algorithms such as a discriminant analysis.

DISCUSSION

The results demonstrate that artificial neural networks outperform classical methods of statistical judgment formation with respect to classification rate, magnitude of the validity coefficient and separability of correctly and incorrectly classified pilot applicants. Furthermore, the results obtained with an artificial neural network were stable in a jackknife as well as a bootstrap validation. In summary it can be said that these results support the criterion validity of the test battery used in this study. However, with regard to practical applications objections are often raised to the use of artificial neural networks in psychological assessment on the grounds that it involves a “black box”, from which the relevance of the individual predictor variables does not follow (cf. Kinnebrock, 1992; DeTienne, DeTienne & Joshi, 2003). This article has shown, however, that this argument does not apply in such general terms. By comparing models with varying number of predictor variables but otherwise identical network architecture it is possible to calculate at least the incremental validity or the relative variance which the various test variables contribute to the predictive model. The weightings themselves, however, remain difficult to interpret. Nevertheless, in practical applications the predictive model described in this article enables empirically validated prediction of pilot applicants’ success in a standardized flight simulator at a rather high level of accuracy and can be used to reduce training costs by selecting the most promising candidates for further military pilot training. However, the authors

acknowledge that the results reported in this paper should be cross-validated using a different sample of military or even civil pilot applicants from various countries in order to further investigate the generalizability of the results.

REFERENCES

- Ajzen, I. (1987). Attitudes, traits and actions: Dispositional prediction of behavior in personality and social psychology. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 20, pp. 1-63). New York: Academic Press.
- Arendasy, M., Hornke, L.-F., Sommer, M., Häusler, J., Wagner-Menghin, M., Gittler, G., Bogner, B., & Wenzl, M. (2005). Manual Intelligence-Structure-Battery (INSBAT). Mödling: Schuhfried GmbH.
- Biehl, B. (1996). Manual Tachistoscopic Traffic Perception Test (TAVTMB). Mödling: Schuhfried GmbH.
- Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford: Oxford University Press.
- Bortz, J. (1999). Statistik für Sozialwissenschaftler [Statistics for social scientists]. Berlin: Springer.
- Brown, M. T., & Wicker, L. R. (2000). Discriminant analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), Handbook of applied multivariate statistics and mathematical modeling (pp.209-234). San Diego, CA: Academic Press.
- Burke, E., Hobson, C. & Linksy, C. (1997). Large sample validations of three general predictors of pilot training success. International Journal of Aviation Psychology, 7, 225-234.
- DeTienne, K. B., DeTienne, D. H. & Joshi, S. A. (2003). Neural networks as statistical tools for business researchers. Organisational Research Methods, 6, 236-265.
- Fahlman, S. E. (1988). Faster-learning variations on back-propagation: an empirical study. Proceedings of the Connectionist Models Summer School. Los Altos: Morgan-Kaufmann.
- Gittler, G. (1998). Manual Adaptive Three-Dimensional Cube Test (A3DW). Mödling: Schuhfried GmbH.
- Goeters, K.-M. (1998). General standards of selection: Validity and utility analysis. In K.-M. Goeters (Ed.), Aviation Psychology: A science and a profession (pp.103-112). Aldershot: Ashgate.
- Griffin, R. B. (1998). Predicting Naval Aviator Flight training Performance using multiple regression and artificial neural networks. International Journal of Aviation Psychology, 8, 121-135.
- Häusler, J. (2004). Software NN Predict. Vienna: self-published
- Häusler, J. & Sommer, M. (2006). Neuronale Netze: Nichtlineare Methoden der statistischen

Urteilsbildung in der psychologischen Eignungsdiagnostik [Neural networks: Non-linear methods of statistical judgment formation in personnel selection]. *Zeitschrift für Personalpsychologie*, 5 (1), 4-15.

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.

Kinnebrock, W. (1992). Neuronale Netze [Neural networks]. München: Oldenburg Verlag.

Lienert, G. A. & Ratz, U. (1998). Testaufbau und Testanalyse (6. Auflage) [Test construction and application (6th edition)]. Weinheim: Psychologie Verlags Union.

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

Mielke, A. (2001). Neuronale Netze [Neural networks]. [Online] URL: <http://www.andreas-mielke.de/nn.html> [01.10.2001].

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

Ree, M. J. & Carretta, T. R. (1996). Central role of g in military pilot selection. *International Journal of Aviation Psychology*, 6, 111-123.

Rojas, R. (2000). Neuronal Networks. A systematic introduction. Heidelberg: Springer.

Wagner, M. & Karner, T. (2003). Manual Cognitrone (COG). Mödling: Schuhfried GmbH.

Warner, B., & Misra, M. (1996). Understanding neural networks as statistical tools. *The American Statistician*, 50, 284-293.

Wittmann, W., & Süß, H.-M. (1997). Challenging G-mania in intelligence research: answers not given, due to questions not asked. Paper presented at the International Meeting of the International Society for the study of individual differences, 19-23 July, Aarhus, Denmark.