Identifying the aviator:

Predictive validity of the selection tests of the Royal Netherlands Air Force.

Author: Suzanne M.A. van Trijp, BSc.
Mentors: prof. dr. Willem B. Verwey, drs. Sebie J. Oosterloo, and drs. Ralph M. Tier
University of Twente, The Netherlands
Royal Netherlands Air Force

This validation study was conducted by Suzanne M.A. van Trijp, BSc. to fulfil a Master's degree in Psychology at the University of Twente, The Netherlands. The validation study was conducted in cooperation with the Royal Netherlands Air Force at its Centre for Man in Aviation, Soesterberg, The Netherlands. Mentors were: prof. dr. Willem B. Verwey and drs. Sebie J. Oosterloo for the University of Twente, and drs. Ralph M. Tier for the Royal Netherlands Air Force.

I wish to thank Willem Verwey and Sebie Oosterloo from the University of Twente. Next to this, I want to thank Ralph Tier for his guidance on my 'endeavour' and correction of many type-os. Dear Calvin, I hate you too! Love Suzy. Bengel, thank you for giving me much needed time. Nico, thank you for the lovely weekends in between, taking care of our cat, and the immensely high phone bills.

Correspondence address: S.M.A. van Trijp, Regulusstraat 11, 7521 DW, Enschede.
E-mail: suzanne@leonerode.nl

"A successful pilot is a high-spirited, happy-go-lucky sportsman who seldom takes his work seriously but looks upon 'Hun-strafing' as a great game and returns after a day's flying to the theatre, music, dancing, and cards."

(Rippon & Manuel, 1918)

"Quiet, methodical men are among the best flyers..."

(Dockeray & Isaacs, 1921)

Front page picture: Defensie beeldbank (2008).

Abstract

A validation study on the selection tests of the Royal Netherlands Air Force was performed by the University of Twente, The Netherlands in cooperation with the Royal Dutch Airforce (RNLAF). This validation study was performed according the research question: What is the predictive validity of the selection tests of the RNLAF concerning the chances of passing/failing the Elementary Military Flight Training (EMFT)? The selection tests that were analysed were the tests of two psychological assessments and two job sample tests. The psychological assessment tests were formed by an instrument interpretation test, a sensori motor coordination test, a dichotic listening test, and six personality competencies based on an interview, personality tests and group assignments. The job sample tests consist of a set of automated (simulator) flight and a set of real flights. Predicting whether a trainee in the selection tests would be able to pass the EMFT is called classification. A need for knowledge on classification errors lead to hypothesis 1: Using the predictors of the selection tests of the Royal Netherlands Air Force causes a change in wrong classification when compared to classification without predictors. Findings in previous research lead to hypothesis 2 and 3. Hypothesis 2: The capacities measured in the first psychological assessment are the predictors with the greatest influence on the probability of correctly classifying the pass/fail EMFT criterion? Whereas hypothesis 3 is: The scores measured in the simulator flights, and scores measured in the real flights are the predictors with the greatest influence on correctly classifying the pass/fail EMFT criterion. Whethers predictor also add predictive value independently was hypothesis 4.

Data was used from digital and paper dossiers and consisted of obtained scores on selection tests obtained by trainees that had succeeded all selection tests, and participated in the EMFT, thus both failed and passed. The sample consisted of 110 cases of trainees that participated in the EMFT between 2005 and 2008. The sample had a passing rate of 56.4%, n= 62. Predictors were chosen based on interviews and kept mostly at end scores of tests. A backward logistic regression analysis was performed with passing/failing EMFT as criterion. Predictors were transformed to standardised Z-scores. Results from analysis were compared to results from a base model. This model contains a constant but does not include any predictors.

The model produced by the analysis was reached in twenty steps and contained the predictor mental load in the real flights. This model showed an overall correct classification of 61.1%; 40.7% positives; 20.4% negatives; 25.9% false positives, and 13.0% false negatives. This supported hypothesis 3 partly. The analysis of group and individual predictors showed that predictors from the real flights were significantly predictive of passing/failing the EMFT, this provided support for hypothesis 4. Analysis of a full model including all predictors showed a 75.9% overall correct prediction and one significant predictor being the mental load of the real flights. Classification results changed due to use of predictors compared to the base model giving support for hypothesis 1. Hypothesis 2 could not be supported.

Contents

Acknowledgements 3
Abstract 5
Contents 7
List of abbreviations 9

1. A validation study on the selection tests of the Royal Netherlands Air Force 11
- 1.1 Selecting and training military aviator 11
- 1.2 Research into aviator selection 12
- 1.3 The selection tests of the Royal Netherlands Air Force 15
- 1.4 Research question and hypotheses 19

2. Data collection and dataset 23
- 2.1 Gathering the data 23

3. Research methods 25
- 3.1 Sample description 25
- 3.2 Test administration: apparatus and method 25
- 3.3 Pre-analysis 26
- 3.4 Predictors and criterion description 26
- 3.5 Statistical Analysis 28

4. Results 31
- 4.1 Sample description 31
- 4.2 Predictors and base model 31
- 4.3 Backward logistic regression analysis 33
- 4.4 Forward logistic regression analysis 36
- 4.5 Added predictive value of groups and individual predictors 36

5. Discussion and conclusion 41
- 5.1 Research question 41
- 5.2 Hypothesis 1 41
- 5.3 Hypothesis 2 and 3 42
- 5.4 Hypothesis 4 43

6. Recommendations 45
- 6.1 Future research recommendations 45
- 6.2 Practical recommendations for the RNLAF 46

7. References 49

Appendix A: Regression Analyses 53

List of Abbreviations

APSS	Automated Pilot Selection System
CMA	Centre for Man in Aviation
EMFT	Elementary Military Flight Training
NLDA	Netherlands Defence Academy
PFS	Practical Flight Selection
RLNAF	Royal Netherlands Air Force

1. A validation study on the selection tests of the Royal Netherlands Air Force

1.1 Introduction

Ask any young child, what they want to be when they grow up and chances are that they answer they would like to be an aviator. The road to becoming an aviator is long; consisting of selection tests, military training, and flight training. A career as a military aviator is only for the few. The aviator selection involves a thorough procedure. When the selection procedure is sound the best candidates are selected. The Royal Netherlands Air Force [Koninklijke Luchtmacht] (RNLAF) wishes to uphold the quality of the selection procedure and therefore gave the assignment to conduct a validation study. Before describing the validation study a general sketch of the selection procedure and information on general aviator selection is given. A detailed description is given in paragraph 1.3.

The first step in the selection procedure is an aviator information day. During this day applicants attend presentations and are able to ask questions to the crew about their working lives and experiences. The day ends with a demonstration flight (RNLAF [1], 2008).

The second step contains the selection tests of the RNLAF. These tests are discussed in detail in paragraph 1.3 “The selection tests of the Royal Netherlands Air Force”. Generally, selection tests where aptitudes, abilities, and skills are measured are the biggest hurdle in the selection procedure (RNLAF [2], 2004).

After completing the selection tests applicants attend the Netherlands Defence Academy [Nederlandse Defensie Academie] (NLDA) where an initial military training is offered that prepares applicants to be officers. Basic and advanced military skills are taught in a period from six months to a year (RNLAF [2], 2006).

Once basic and advanced military skills are mastered, the officers/trainee aviators transfer to the Elementary Military Flight Training [Elementaire Militaire Vlieger Opleiding] (EMFT). The trainee aviators in the EMFT complete ground school (theory of flight) followed by flight training in the Pilatus PC-7(RNLAF [2], 2006).

A solo flight completes the EMFT, after which trainee aviators are appointed to fixed wing or rotary wing according to their performances and numbers of places available in the additional flight training. Those who are top of their class are selected for fixed wing; the others are selected for rotary wing. Trainee aviators continue their education in the United States of America where they receive additional flight training and type specific flight training[1]. The duration of additional flight training and type specific training is approximately one year, after which the trainee aviator receives a wing[2] (RNLAF [2], 2006).

Back in the Netherlands aviators follow a conversion training aimed at flying in the Dutch climate and circumstances. After completing this training the aviators are placed at a squadron and start their operational career (RNLAF [2], 2006).

1.2 Research into aviator selection

1.2.1 History and measures

At first, military aviator selection was developed in Italy in the period prior to the First World War and measured reaction time, emotional reaction, equilibrium, perception of muscular effort, and attention. During the First World War more countries applied selections to reduce the high attrition rate in the aviator training. This attrition rate could be up to 90% (Hunter & Burke, 1995). Measures of intelligence seemed effective. The interbellum was characterized by a growth in selection research in the United States of America and Germany (Hunter & Burke, 1995). The American Army Air Corps put the focus on measuring general mental and reasoning abilities. The German Air Force focused mainly on subjective measures with tests such as Rorschach (Tsang & Vidulich, 2008). During the Second World War there was renewed interest in selection research stretching the topics of selection to: intelligence, psychomotor skill,

[1] Type specific training for fixed wing: Cessna T37 Tweet, T38, and F16 Fighting Falcon. For rotary wing: TH67 creek, Huey, Cougar, Chinook, and AH-64 Apache.

[2] The ‘wing’ is a brass set of miniature wings that can be placed on a uniform to indicate that the person is an aviator. This decoration is highly valued and desired within the RNLAF.

mechanical comprehension, and spatial measures. After the Second World War testing of personality became important. From the 1970's to present day all aviator selections test multiple aptitudes and psychomotor abilities (Tsang & Vidulich, 2008). In addition, personality measurements are common in continental Europe (Hunter & Burke, 1995).

1.2.2 Previous validity research

Many validation studies on military aviator selection tests have been undertaken (Martinussen & Torjussen, 1998., Delaney 1992). Often due to small samples sizes, small variances, range restriction, and dichotomization results were neither staggering nor significant. In general, it seems that a general cognitive factor 'g' has the best predictive validity, especially when this general cognitive factor is tested together with other constructs (Tsang & Vidulich, 2008, Hunter & Burke, 1995).

In 1997, Burke, Hobson, and Linsky performed a meta-analysis in which a composite data file of several data files from different air forces was used for analysis. This ensured a large sample. Constructs tested in all air force selections were chosen as predictors. They examined predictive validity of: control of velocity, instrument interpretation, and sensori motor apparatus. The criterion was pass/fail flight training score. Conclusions were that the composite observed validity was r=.24 without any corrections.

Martinussen and Torjussen (1998) found that the predictive validity of the Norwegian test battery on criteria of basic military flight training was high for an instrument interpretation test (r= .29), a mechanical principles test (r=.23), and aviation information (r= .22).

Delaney (1992) conducted a validation study in which the predictive validity of a dichotic listening task and a psychomotor task on primary flight training criteria were tested. This study showed that a combination of performance scores on the dichotic listening task and the psychomotor task show a multiple regression coefficient of R=.442. Individual results were: psychomotor test r=.26 to .44 and dichotic listening task r= .22 to .28. Hunter and Burke (1995) [2] further summarized that many studies showed a correlation between actual flying and job sample tests such as simulator based flying. Job sample tests were described as: *“an artificially created situation in which an individual is required to perform either the same tasks that will be performed on the job, or tasks that are very similar to those that will be performed on the job.”* (Hunter and Burke, 1995).

Recently the Portuguese Air Force presented a study in which they compared several classification methods to predict flight success in military pilots (Marques & Gomes, 2008). Though its goal was to compare classification methods some predictive results also surfaced. With a sample of 254 aviators they tested the predictive validity of 10 predictors on a pass/fail criterion in the flight screening, which is the fourth phase of Portuguese Air Force selection. Neural networks analysis, discriminant analysis and logistic regression showed that predictors were instrument interpretations test 1 and 2 (information processing and spatial aptitude), sensorimotor apparatus (sensomotor coordination), and vigilance (attention).

1.2.3 Previous validity research of the RNLAF

Research conducted by the RNLAF in 2005 (RNLAF [3], 2005) focused mainly on predictive value of flying aptitude tests on the Elementary Military Flight Training (EMFT). The job sample test scores Automated Pilot Selection System (APSS) and Practical Flight Selection (PFS) were analysed against the pass/fail criterion of the EMFT. Capacity and personality tests were a priori excluded. Participants of this research joined the EMFT from 2000 to 2005 and therefore this research is a direct predecessor of the current validation study. With n=122 and a pass rate of 66% it was found that from the APSS the best predictors were the flight score of the last flight and the mental load scores of the second and third flight. With these predictors 79% of all participants’ passing or failing was predicted correctly. For the PFS it was found that the fourth flight was a good predictor that ensured correct classification in 77% of all the cases.

1.2.4 Conclusions

The RNLAF selection tests do not include all discussed tests. Tests measured in other research that the RNLAF uses as well are: instrument interpretation, sensori motor apparatus, dichotic listening task, and job sample tests. Results from previous research indicate that highest predictive validity can be expected in this validation study from all above noted tests. Personality tests have not been taken into account in previous research and any results in this area are new. The general cognitive factor g has been shown to predict well. However, it is not tested by the RNLAF in its selection tests and cannot be taken into account in this validation study.

1.3 The selection tests of the Royal Netherlands Air Force

In this paragraph all the selection tests of the RNLAF will be presented and discussed in detail. Variables and procedures will be explained for each test divided over several subparagraphs. The first subparagraph contains general information about the selection procedure. After this, separate selection rounds will be described.

1.3.1 General information on the selection procedure of the RNLAF

As sketched in paragraph 1.1 aviator applicants have to complete a selection procedure prior to being appointed as an aviator. Applicants can either be external applicants, or employees of the RNLAF who wish to apply for an aviator (related) position.

The selection procedure starts with an administrative pre-test and ends with a medical examination (Tactische Luchtvaart [Tactical Air Force], 2007). The administration and medical part of the application process are not in the scope of this study. Selection tests are the scope of this study.

The selection tests are divided into four separate stages that take place at the Centre for Man in Aviation [Centrum voor Mens en Luchtvaart] (CMA). Tests are conducted by psychologists and assistant psychologists, who work by rules and standards, set by the Netherlands institute for psychologists [Nederlands instituut voor psychologen] to ensure professional ethics. In the selection procedure an up-or-out system is followed. When the applicant fails in a certain stage the application is either put on hold for a period of time or the application is terminated. When the applicant passes a stage, he or she goes on to the next stage. The four selection stages are: first psychological assessment, automated pilot selection system, second psychological assessment, and practical flight selection. Norms, standards and methods of the selection tests have changed substantially around 2005. After 2005 the tests largely remained the same (Tactische Luchtvaart [Tactical Air Force], 2007).

1.3.2 The first psychological assessment

The first psychological assessment consists of three separate tests.

1. In the instrument interpretation test, applicants combine information from a compass and an altitude device and then select the correctly depicted airplane out of several options. The goal of the instrument interpretation test is to measure spatial aptitude (RNLAF [4], year unknown).
2. In the sensori motor coordination test, applicants must keep a continuously hovering form on a specific spot using a joystick and foot pedals. This test measures sensomotor skills (Parker, G. and Oliver, N. 2006)
3. In the dichotic listening task, applicants have to discriminate the correct message from two offered messages, each on one ear, while being primed to one of both ears. The dichotic listening task measures the applicants' ability for attention switching (RNLAF [5], year unknown).

Applicants who pass the first psychological assessment are allowed to go on to the next stage: the automated pilot selection system. When applicants fail on one of the tests in the first psychological assessment, their application is put on hold for a period of six months, after which a second chance is offered (A. Lablans, personal communication, May, 06, 2008).

1.3.3 The Automated Pilot Selection System

The next stage in the selection procedure consists of the APSS, in which at least three and a maximum of five simulated flights with an increasing level of difficulty are flown. The theory of simulated flying is studied by the applicant beforehand, study material is provided by the RNLAF. The simulated flight tests measure flying aptitude.

Performance on the first three flights determines whether an applicant is allowed to fly the last two flights. When results show that an applicant performs below standards, the application is terminated after the third flight and the applicant cannot apply ever again. Applicants who are allowed to fly the last two flights are assessed after completing these flights. Those who perform up to mark may go on to the second psychological assessment. For those who do not pass the simulated flights the application is permanently terminated. Exceptions to an application termination rarely occur (W.A.C. Helsdingen, personal communication, April, 23, 2008).

1.3.4 The second psychological assessment

The second psychological assessment focuses on competencies and the applicant's motivation. Applicants fill in four personality questionnaires and they participate in several group assignments during which their behaviour is observed. To complete the assessment the applicant is interviewed by a psychologist.

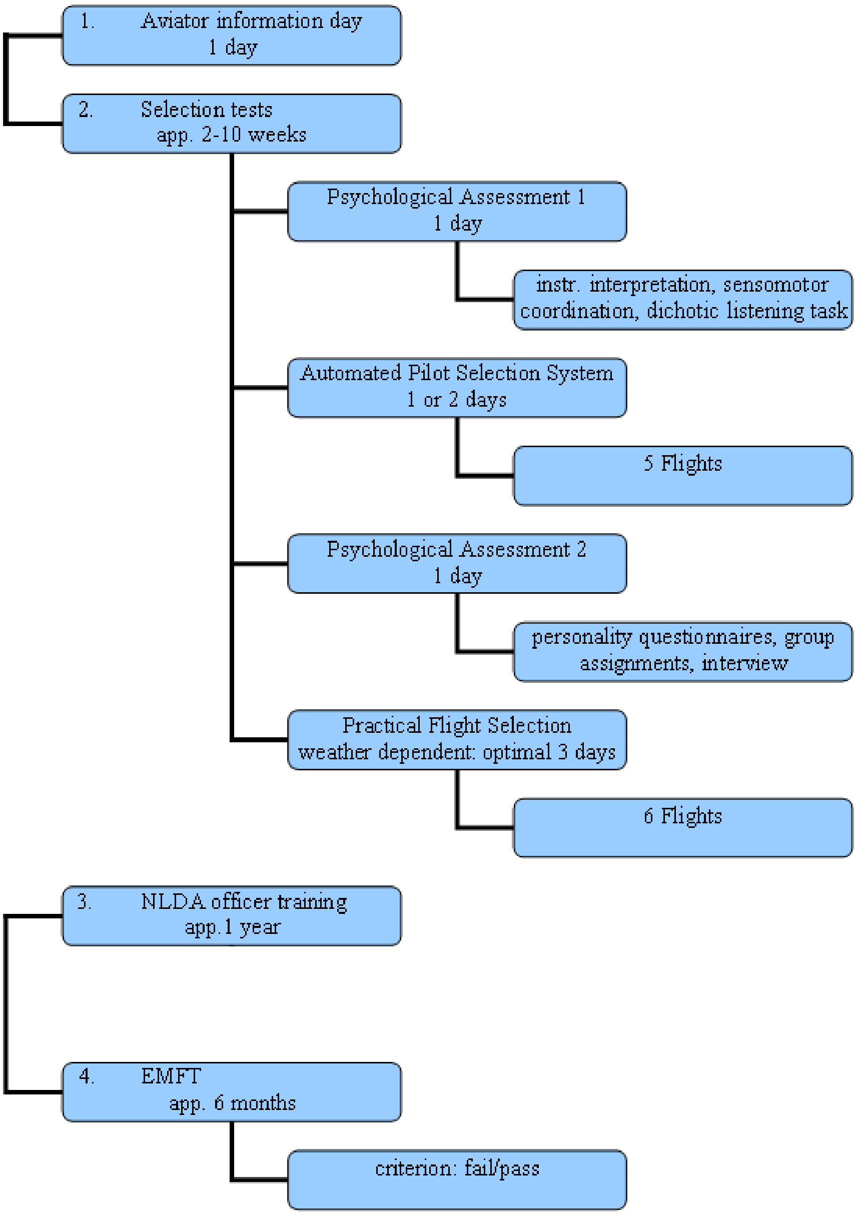
The application of applicants that fail the second psychological assessment is put on a temporary hold. Applicants can redo their application from the second psychological assessment on, either after a period of one year, or in special occasions after a period of six months (R.M. Tier & A.C. van Beersum, personal communication, May, 07, 2008).

1.3.5 The Practical Flight Selection

The last hurdle in the selection procedure is the PFS. A maximum of six practical flights with increasing difficulty are offered to the applicant. The first flight is a familiarization flight and an indicator of airsickness. Since 2008 the PFS takes place in Portugal. Before 2008 the PFS took place in Seppe, The Netherlands. A clear sky is more likely in Portugal than in The Netherlands. This is important since good visibility of the horizon when flying the PFS is a must. Applicants are judged on flight aptitude, mental load and their progression.

Applicants that pass the PFS go on to a medical examination and receive a graded application advice. These grades are: excellent, good, or average (Tactische Luchtvaart [Tactical Air Force], 2007). Those who fail the PFS see their application terminated permanently. An alternative is offered to apply for the position of air combat controller (F. Jurres, E. Jurres & C.M. van Nieuwburg, personal communication, May, 19, 2008).

An overview of the selection procedure, its tests, approximate duration, and initial training can be found in Figure 1.



1. Aviator information day
1 day
2. Selection tests
app. 2-10 weeks
Psychological Assessment 1
1 day
instr. interpretation, sensomotor coordination, dichotic listening task
Automated Pilot Selection System
1 or 2 days
5 Flights
Psychological Assessment 2
1 day
personality questionnaires, group assignments, interview
Practical Flight Selection
weather dependent: optimal 3 days
6 Flights
3. NLDA officer training
app.1 year
4. EMFT
app. 6 months
criterion: fail/pass

Fig. 1. Overview of the selection (number 1 and 2 and their branches) and the first part of training of the RNLAF (number 3 and number 4 and its branche). The branches of number 1 and 2 are connected since they belong to the application tract of the aviator applicant. Branches 3 and 4 are separate from the application tract since applicants are hired by the RNLAF in these stages.

1.4 Research question and hypotheses

The goal of the RNLAF selection procedure is to select the ideal candidates to be trained as military aviators. To meet this goal the selection procedure must have a high predictive validity and must measure constructs that are highly predictive of the performance of trainee aviators. Since 2005 no validation study has been performed. Therefore it is unknown what the predictive validity of the RNLAF selection procedure (Fig 1, part 2. Selection tests on previous page) for the fail/pass criterion in the EMFT is for the period of 2005 to 2008. Norms, standards and methods of the selection tests have changed substantially around 2005. Therefore scores from tests taken before 2005 are not included in this validation study. After 2005 the tests largely remained the same (Tactische Luchtvaart [Tactical Air Force], 2007).

This leads to the following research question: *What is the predictive validity of the pilot selection tests (the tests of psychological assessments 1 & 2, the automated pilot selection system, and the practical flight selection) of the Royal Netherlands Air Force concerning the chances of succeeding the Elementary Military Flight Training for the years 2005 to 2008?*

1.4.1. Statistical testing

For the RNLAF it is important to keep the number of persons that fail the EMFT when they were predicted to pass as low as possible. In statistical terms these persons are called: false positives. The persons that are predicted to fail but would pass if they were to take part in the EMFT are called false negatives. A high percentage of false positives would cost the RNLAF money while a high percentage of false negatives would cause the RNLAF to miss out on potentially good aviators.

When looking at the predictive validity of selection tests it is thus also important to address the change of both false positives and false negatives, also known as a change in wrong classifications. This leads to the following hypothesis:

Hypothesis 1: Using the predictors from the selection tests of the RNLAF causes an change in wrong classification when compared to classification without predictors.

Conclusions drawn from previous validity research lead to several hypotheses. Flight aptitude was highly correlated with several capacities (information processing, spatial aptitude, sensomotor skills, and vigilance) (Marques and Gomes, 2008; Burke, Hobson, & Linsky, 1997; Martinussen and Torjussen,1998; Delaney,1992);, therefore it is hypothesized that scores of the first psychological assessment are better predictors of pass/failing the EMFT than other selection test scores. This effect is displayed when a raise of obtained scores of the first psychological assessment has a greater effect on the chances of passing/failing the EMFT than when obtained scores of other selection tests are raised. This hypothesis is:

Hypothesis 2: The capacities measured in the first psychological assessment are the predictors with the greatest influence on the probability of correctly classifying the pass/fail EMFT criterion.

Next to capacity tests, job samples were found to be highly predictive for the chances of passing/failing initial military flight training (RNLAF [3], 2005; Hunter and Burke, 1995). The RNLAF's job sample test results are partly definitive in the selection procedure in the sense that a negative result means candidates are excluded from application forever.

This procedure suggests that not the scores of the first psychological assessment but scores of the APSS and PFS are the better predictors of chances of passing/failing the EMFT. This effect would be shown when a raise of obtained scores of the APSS and PFS has a greater effect on the chances of passing/failing the EMFT than when obtained scores of other selection tests are raised. This hypothesis is:

Hypothesis 3: The scores measured in the automated pilot selection system, and the scores measured in the practical flight selection are the predictors with the greatest influence on the probability of correctly classifying the pass/fail EMFT criterion.

Lastly, it is important to know whether predictors or sets of predictors add predictive value to a model, when they are analysed independently instead of all predictors together in a model, or not. This leads to a final hypothesis:

Hypothesis 4: Individual predictors or sets of predictors add predictive value to the base regression model.

2. Data collection and dataset

2.1 Gathering the data

2.1.1 RNLAF data archives

The data set used in this research comprises several data subsets. These subsets are: scores of the psychological assessment 1 and 2, scores of the Automated Pilot Selection System (APSS), scores of the Practical Flight Selection (PFS), and scores of the criterion fail/pass in the Elementary Military Flight Training (EMFT). A paper file of each applicant is kept at the Centre for Man in Aviation (CMA), Soesterberg, The Netherlands, with all his or her scores collected. The different selection departments keep a separate digital archive as well. Digital scores of assessments and scores of APSS are at the CMA. The digital PFS scores are kept in Seppe, The Netherlands. Criterion scores of failed trainee aviators are kept in the primary military flight school; scores of passed trainee aviators are added in the personal logs of aviators.

2.1.2 Data problems

Several problems occurred in the data gathering process.

Firstly, the data subsets were not archived in a central place. Even though selection scores are kept together in an applicant's file, digital data can only be retrieved from separate databases by assigned personnel. The downside of this approach is that the dataset is fragmented; it takes longer to reconstruct and the resulting dataset needs to be crosschecked to make sure it is complete and correct.

Secondly, the APSS scores were not available in a digital format causing extra workload; it took one month to assemble. Digital databases are far more efficient in use.

Thirdly, the company that performs the PFS needed digital scores of the APSS to be able to find requested data in their digital archives. Therefore data of PFS scores could only be retrieved after APSS scores were digitalised.

Fourthly, the primary flight school does not keep a record of their input and output; it does not provide data on pass/fail results or provides lists of trainee aviators starting the EMFT. The lack of data caused a time delay. Next, it induced piecing together fragments of data from different sources. This is prone to errors, time-consuming, and implies that different persons need to be given approval of access, increasing the chance of delay.

Fifthly, PFS scores were not easily available because they were stored at an external company and because this company was eventually not willing to compose a database with scores to be used for present study scores seemed not available at all. Most PFS scores then needed to be completed manually via the personnel dossiers stored at the CMA. Some personnel dossiers were missing causing extra missing values and a time delay.

Besides problems in data gathering there was also a gap in the database itself due to a crashed computer network in the past. Scores of the second psychological assessment for the period 2005 were lost. This needed manual reconstruction of 110 cases based on paper files.

The incompatibility of the data formats posed another problem. Though software can import and export numbers between SPSS and Microsoft Excel a part of the data information is lost. The numbers are imported, however variable information behind data is lost. This is problematic since names of variables and labels within variables are lost. Completing this for one or two variables is straight forward but completing this for 20 variables takes up time and is prone to errors.

When examining the data another problem came to light. Scores of the sensori motor test could not be found. Instead there were scores of a previously used sensori motor test. Since the measured constructs are alike in both tests the scores of a previously used test can be used (C.M. van Nieuwburg, personal communication, May, 2008).

3. Research methods

3.1 Sample description

Descriptive statistics were calculated on the independent variables: 'Gender' and 'Age'. The research sample constitutes of trainee pilots that have passed all selection tests, passed officers training and participated in primary military flight school from classes 2005 to the last class of 2007. Participants are those that entered the EMFT and either passed or failed the EMFT. The sample consists of N=110 cases. Pass rate of the EMFT for classes 2005 to 2007 is: 56.4%, n=62.

3.2 Tests administration: apparatus and method

3.2.1 The first psychological assessment

All tests of the first psychological assessment (instrument interpretation, sensomotor coordination and dichotic listening task) are administered on a PC, one per applicant in a large classroom. The instrument interpretation and dichotic listening task are administered via a regular keyboard. The sensori motor test however, is tested via a specially designed console and a set of foot pedals.

3.2.2 APSS

The automated flights can be administered on three different types of simulators. The differences that appear in flight difficulty because of these different simulators are corrected for by the computer to make sure output scores are comparable. Applicants are tested individually by an instructor with an instructor change after three flights. There are pre-flight and post-flight briefings. After three flights a lunch break is included.

3.2.3 The second psychological assessment

Applicants undertake four personality tests on a PC. Secondly, applicants take part in a series of group assignments with obtrusive observation. Thirdly, applicants will have an individual interview with a psychologist.

3.2.4 PFS

The PFS takes place in a Slingsby T-67 Firefly; see Figure 3 for an example of the aircraft.

Fig. 3. Slingsby Firefly at TTC Seppe, the Netherlands. Photographer: A. Vercruijsse

During the test the instructor is seated alongside the applicant. Duration of the PFS is depending on weather conditions but lasts a minimum of two days to give the applicant the chance to recuperate between two sets of three flights.

3.3 Pre-analysis

The first step was to identify applicants in the raw dataset that have succeeded all selection tests, succeeded officers training and participated in the EMFT. This happened in a retrograde way.

The next step was to choose the predictors used in the analyses. This was done by choosing predictors that reflected end scores or summary scores. A detailed explanation on the choice of predictors can be found in paragraph 3.4. Lastly, all the cases in the research were coded for privacy protection.

3.4 Predictors and criterion description

3.4.1 Predictors and criterion

Predictors used in the validation study were derived directly from the selection tests of the RNLAF. An overview of independent predictors can be found in Figure 4. The criterion used in this research was the Elementary Military Flight Training (EMFT): pass or fail. This criterion is dichotomous.

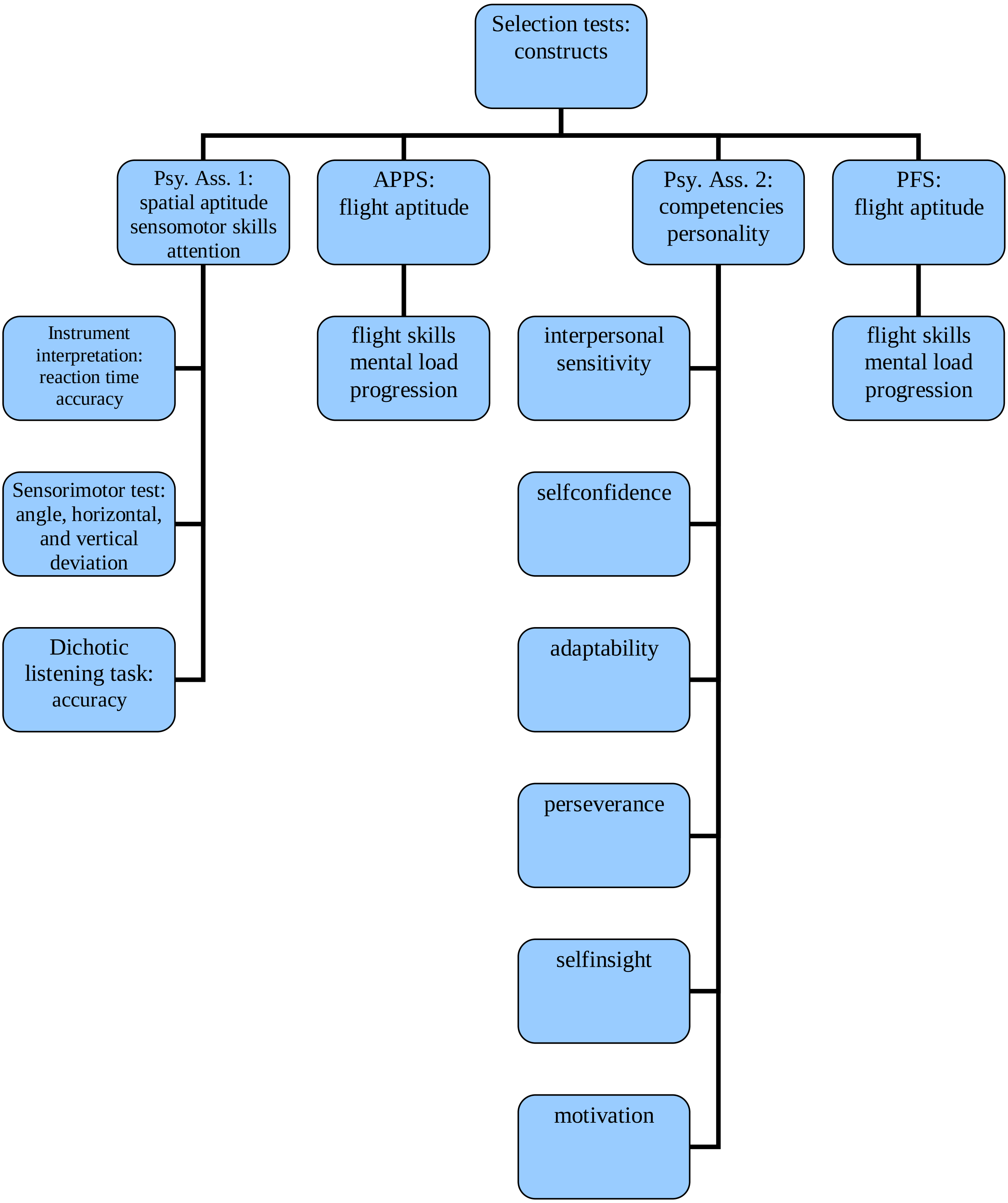

Selection tests: constructs
Psy. Ass. 1: spatial aptitude sensomotor skills attention
APPS: flight aptitude
Psy. Ass. 2: competencies personality
PFS: flight aptitude
Instrument interpretation: reaction time accuracy
Sensorimotor test: angle, horizontal, and vertical deviation
Dichotic listening task: accuracy
flight skills mental load progression
interpersonal sensitivity
selfconfidence
adaptability
perseverance
selfinsight
motivation
flight skills mental load progression

Fig.4. Predictors used in the validation study. All predictors taken from the selection tests as described in Fig 1. number 2. Selection tests

3.4.2 Argumentation choice of predictors

There were two reasons for the choice of predictors used in this study.

First, in interviews with employees of the department of psychological selection of the RNLAF they advised to use quantitative predictors. They also advised to use end scores of tests.

Second, the basics of regression analysis required an amount of predictors that is small compared to N. When using only end scores and summary scores the number of predictors could be reduced beforehand.

3.5 Statistical analysis

3.5.1 Logistic Regression analysis

To test the predictive validity of the selection tests on the chances of passing/failing the EMFT, a regression analysis was performed. The EMFT criterion (pass/fail) is dichotomous, thus a logistic regression analysis was needed. Stepwise logistic regression is mostly used in explorative research whereas full model regression is often used to test hypotheses. This research sample with a small N and a large amount of predictors called for an explorative approach.

A backwards stepwise logistic regression was chosen[3]. All predictors were placed in a model and those that did not contribute to the criterion were eliminated from the model through a series of steps. At the end of the analysis a model has been build that included predictors that had significant predictive value on the criterion. Each building block in the model carried the predictive value of the predictor on the criterion. A forward stepwise logistic regression was performed as a check.

To show added predictive value of individual predictors or groups of predictors, additional logistic regression analyses were performed where per analysis only one predictor or one group of predictors was analysed against the base model.

[3] Background information on logistic regression and the difference from linear regression analysis can be found in Appendix A.

3.5.3 Restriction of range

Restriction of range effect appears when data are only used from applicants who have met qualifying selection scores. This dataset then contains a small variation. A full range of scores would be available if all applicants regardless of selection tests would participate in the EMFT. Since this is not the case, restriction of range is expected to influence the results. Restriction of range tends to have a downsizing effect on the regression results (Hunter & Schmidt, 1990).

One study speaks of the possibility of correction for restriction of range through artificial extrapolation of extreme data to regular data (Dunbar & Linn, 1991). This correction has not been used.

4. Results

4.1 Sample description

4.1.1 Descriptive statistics

The sample used in this research has an *N*= 110. Nearly all participants are male, namely 98.2 % (*n*= 108). Leaving 1.8 % females (*n*= 2).

The M_{Age} = 19.6 years with $\text{Minimum}_{\text{Age}}$ = 16 years and $\text{Maximum}_{\text{Age}}$ = 28 years. Most candidates applied after finishing secondary school. A distribution of age can be found in Figure 5.

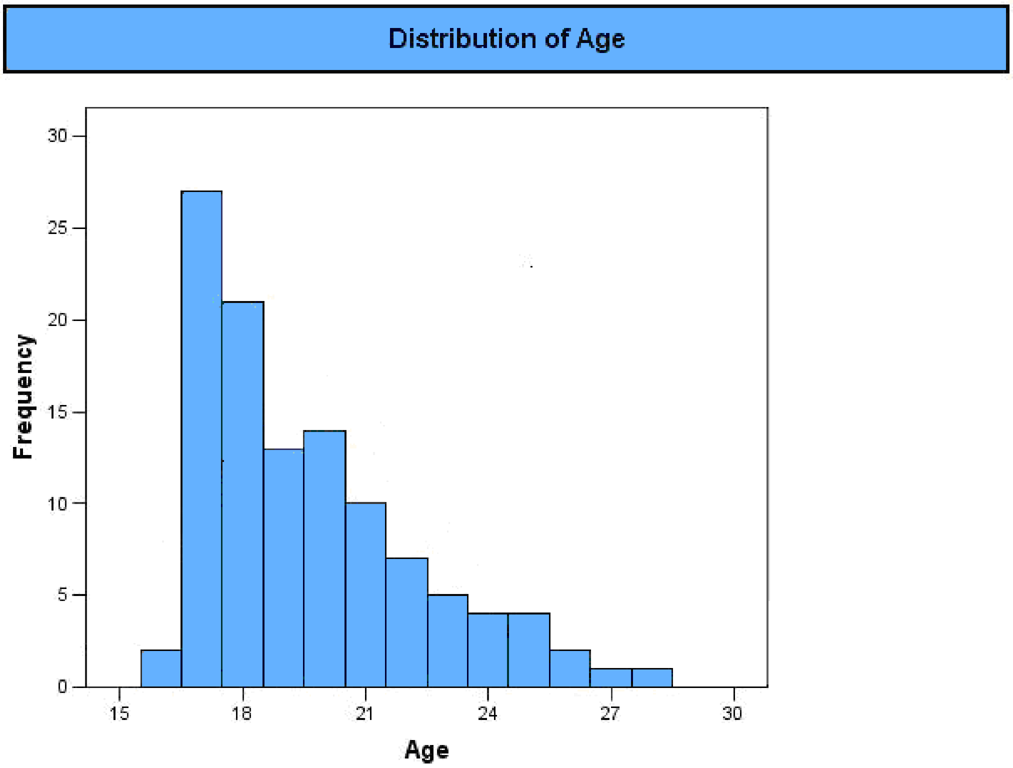


Fig. 5. Distribution of ‘Age’ obtained from sample

4.2 Predictors and base model

4.2.1 Predictors

All predictors were first transformed to standardised Z-scores to ensure that all predictors can be compared with each other. An overview of the predictors and how they are displayed in the constructed model can be found in Table 1.

Selection tests rounds	**Predictors**	**Predictor name in model**
Psychological Assessment 1	instrument interpretation	$\chi_{\text{-ii}}$
	selective listening task	$\chi_{\text{- slt}}$
	sensomotor-coordination	$\chi_{\text{-smc}}$
Automated Pilot Selection System	subtotal flightscore	$\chi_{\text{-sf}}$
	total flightscore	$\chi_{\text{-tf}}$
	endscore overall	$\chi_{\text{-eo}}$
	mental load 1	$\chi_{\text{-m1}}$
	mental load 2	$\chi_{\text{-m2}}$
Psychological Assessment 2	motivation	$\chi_{\text{-mo}}$
	perseverance	$\chi_{\text{-pe}}$
	selfconfidence	$\chi_{\text{-se}}$
	interpersonal sensitivity	$\chi_{\text{-in}}$
	selfinsight	$\chi_{\text{-si}}$
	adaptability	$\chi_{\text{-ad}}$
Practical Flight Selection	subtotal 1 flightscore	$\chi_{\text{-sub1}}$
	subtotal 2 flightscore	$\chi_{\text{-sub2}}$
	endscore flight	$\chi_{\text{-ef}}$
Practical Flight Selection	endscore progression	$\chi_{\text{-ep}}$
	endscore mental load	$\chi_{\text{-em}}$
End ranking grade	ranking grade	$\chi_{\text{-r}}$

Table 1. Overview of predictors used in analyses. All predictors were Z-transformed first.

4.2.2 Model

The RNLAF-selection tests equation consists of all predictors and a constant with a chance (passing/failing the Elementary Military Flight Training (EMFT)) as an end criterion. Equation 1 depicts the RNLAF-selection tests equation:

$$P(y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} \frac{[\exp(a + b_{ii}\cdot c_{ii} + b_{slt}\cdot c_{slt} + b_{\dots}\cdot c_{\dots} + b_r\cdot c_r)]^{y_i}}{1+\exp(a + b_{ii}\cdot c_{ii} + b_{slt}\cdot c_{slt} + b_{\dots}\cdot c_{\dots} + b_r\cdot c_r)} \qquad (1)$$

In addition to the equation, a base model exists. The base model contains a constant and does not include any predictors. Based on the constant the base model can predict classification results. If results indicate that 50% or more is classified as pass then all cases are predicted as passed and thus have a 100%

score of passed and a 0% score of failed. If results indicate that less than 50% is a pass than all cases will be predicted as fail. In the present study the base model results existed of 100% predicted passed, 0% predicted failed and a 53,7% overall correct classification.

4.3 Backward logistic regression analysis

4.3.1 Significant predictors

The backward logistic regression analysis produces a model with significant predictors only. Here, this model was reached in twenty steps and included one predictor. This predictor was χ_{em}; which is the end score mental load of the practical flight selection. In the analysis originally 20 predictors were included. Results are addressed in the discussion and conclusion section.

4.3.2 Classification

The classification results of the model in step 20 showed that using this model increases the percentage of correct predictions by 7.4% when compared to the base model. Furthermore its distributions of false positives went down and false negatives went up compared to the base model. Table 2 shows an example of a classification table. In this table one can see category A, those trainees that were expected to pass and were observed to pass, B those trainees that were predicted to pass but in fact failed (false positives), C those trainees that were predicted to fail but in fact passed (false negatives), and D those trainees that were predicted to fail and indeed failed. Overall correct prediction refers to the percentages A plus D. This classification table format is used for all classification tables from now on.

example classification table			**Observed**	**Observed**	
			EMFT passed	EMFT failed	
model	**Predicted**	EMFT passed	A = positives	B = false positives	All those predicted to pass, A plus B
		EMFT failed	C = false negatives	D = negatives	All those predicted to fail, C plus D
			All those passed, A plus C	All those failed, B plus D	
				model overall correct prediction =	percentage correctly predicted to pass AND to fail

Table. 2. example of classification model

The model showed an overall correct prediction of 61.1%, with a number of 40,7% positives and 20,4% negatives. Part of false positives is 25.9% and a part of 13,0% false negatives.

Table 3 displays classification percentages of the base model and classification results of the model of step 20.

4.3.4 Model and model fit.

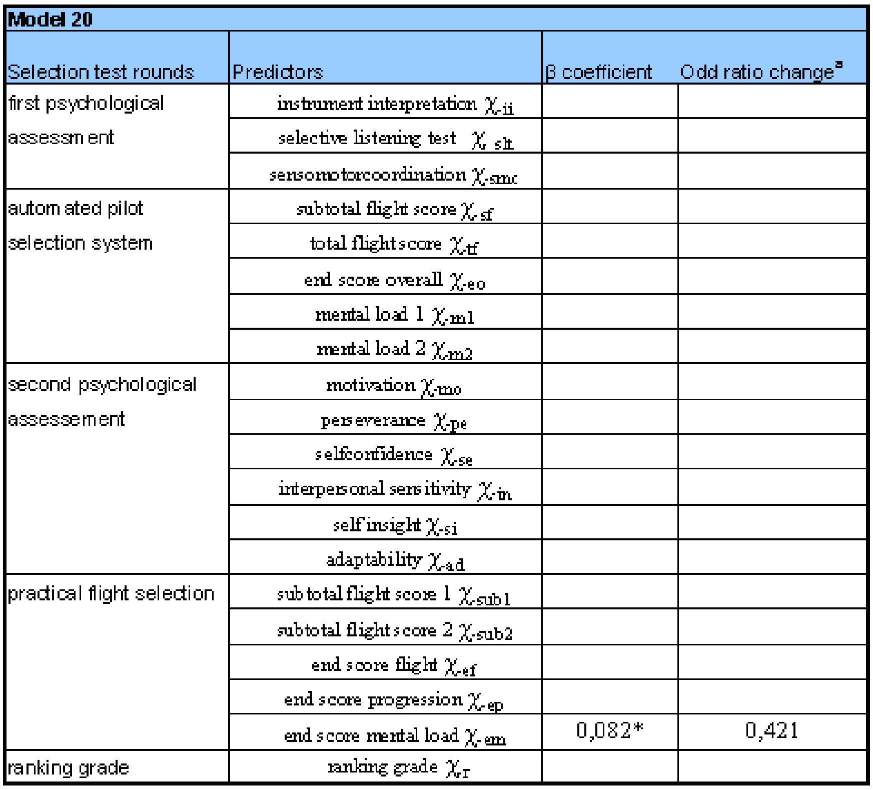
The Hosmer and Lemeshow test gives an indication whether the model describes the population data adequately or not. A poor fit is indicated when $p < 0.05$. The model of step 20 passes the Hosmer and Lemeshow test with a good fit. Step 20: χ^2 (8, N = 110) = 4.638, p = 0.795.

Classification model 20			**Observed**	**Observed**	
			EMFT passed	EMFT failed	
basemodel	**Predicted**	EMFT passed	53,7%	46,3%	100%
		EMFT failed	0,0%	0,0%	0%
			54%	46%	
			basemodel overall correct prediction =		53.7%
			EMFT passed	EMFT failed	
model 20	**Predicted**	EMFT passed	40,7%	25,9%	66,7%
		EMFT failed	13,0%	20,4%	33,3%
			54%	46%	
			model 20 overall correct prediction =		61,1%

Table 3. classification results of base model and classification results of model step 20

For model 20 the pseudo $R^2 = 0.107$ (Nagelkerke). The closer R^2 approaches 1 the more of the variation of the criterion is explained by the model. In this case R^2 is approaching 0 indicating that most of the variation is explained by something else than the model.

The odd ratio change depicts the influence of each predictor on the criterion. In the model of step 20 one predictor is included, namely the end score mental load of the practical flight selection (χ_{em}). The probability of correctly predicting the criterion is proportionally influenced with 0.421 by the χ_{em} predictor. Table 4 gives an overview of the model in step 20 with the β coefficient of the predictor χ_{em}, significance value of the predictor χ_{em}, and the odd ratio change for the predictor χ_{em}. Furthermore, Figure 6 depicts a graph of the probability of predictor $\chi_{em.}$



Model 20			
Selection test rounds	Predictors	β coefficient	Odd ratio change[a]
first psychological assessment	instrument interpretation χ_{ii}		
	selective listening test χ_{slt}		
	sensomotorcoordination χ_{smc}		
automated pilot selection system	subtotal flight score χ_{sf}		
	total flight score χ_{tf}		
	end score overall χ_{eo}		
	mental load 1 χ_{m1}		
	mental load 2 χ_{m2}		
second psychological assessement	motivation χ_{mo}		
	perseverance χ_{pe}		
	selfconfidence χ_{se}		
	interpersonal sensitivity χ_{in}		
	selfinsight χ_{si}		
	adaptability χ_{ad}		
practical flight selection	subtotal flight score 1 χ_{sub1}		
	subtotal flight score 2 χ_{sub2}		
	end score flight χ_{ef}		
	end score progression χ_{ep}		
	end score mental load χ_{em}	0,082*	0,421
ranking grade	ranking grade χ_{r}		

Table 4. Coefficient and odd ratio change of model in step 20. All predictors were Z-transformed. [a] = the ratio change in the odds of the passing/failing EMFT for a one-unit enhancement of a predictor while all others stay equal. * $p < .10$

4.4 Forward logistic regression analysis

Results of the forward logistic regression analysis with $\alpha = 0.10$ show a model that contains the predictor χ_{em}. This concurs with the model from step 20 from the backwards analysis method. Significance and classification results are alike to those in the backwards analysis. The forward analysis method acts as a check on the backwards analysis and in this case validate the backward analysis' results.

4.5 Added predictive value of groups and individual predictors

4.5.1 Groups of predictors

The several groups of predictors analysed are: group first psychological assessment, group automated pilot selection system, group second psychological assessment, and group practical flight selection. These groups were chosen based

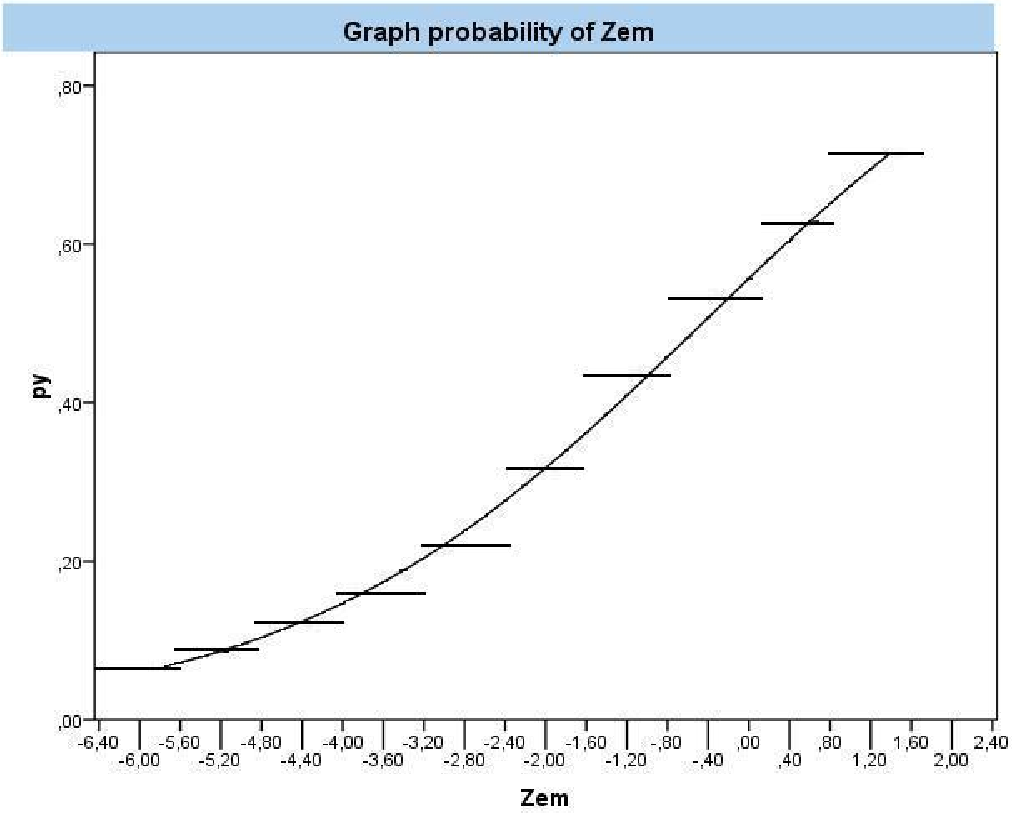


Fig. 6. The probability of predictor $\chi_{em.}$ On the x-axis the scores of χ_{em} are found and on the y-axis the probability of passing the EMFT. Markers are set to indicate fractions of passed trainees and the probability of this fraction

on the selection rounds of the RNLAF. Analyses showed that the group of practical flight selection produces a significant model that increases the predictive value of the base model. This model shows 52,6% positives; 13,4% negatives; 6,2% false positives, and 27,8% false negatives with an overall correct prediction of 66.0%.

This model has a pseudo R^2 of 0.175 (Nagelkerke) and a Hosmer and Lemeshow model fit of p = 0.473 (χ^2 = 13.464, df = 5, p = 0.019**).

4.5.2 Individual predictors

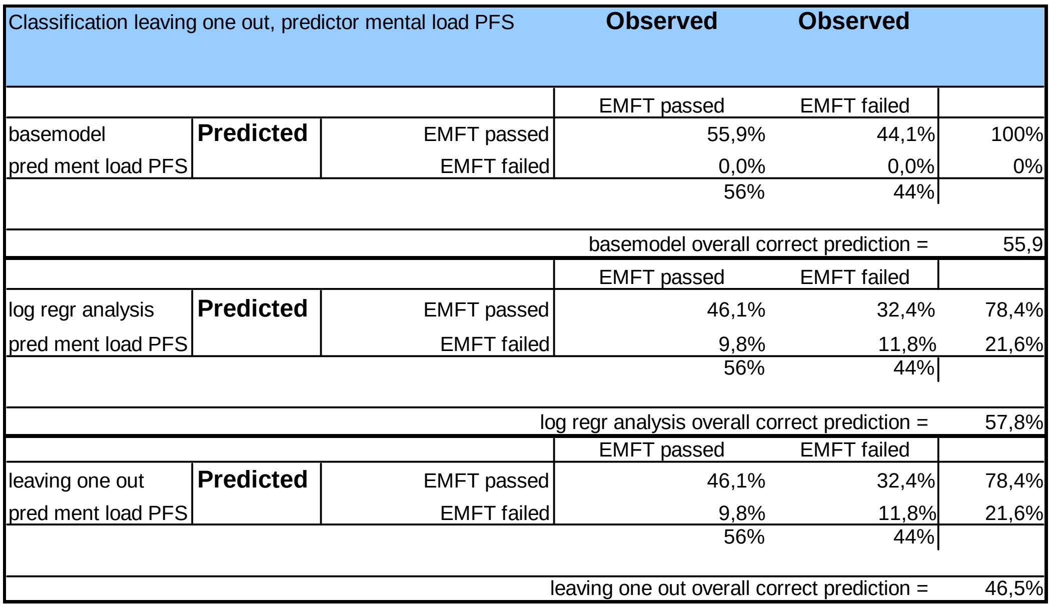
Individual predictors (χ_{ii} to χ_r) were added to the base model and then analysed. Two predictors showed significant added predictive value on the criterion (pass/fail EMFT). This means that these two predictors individually, thus without cooperation of other predictors, have enough predictive value on the criterion to be significant.

The first predictor is χ_{ep}; end progression score of the practical flight selection (β = -0.390, p = 0.067*). The classification results of the model with χ_{ep} as predictor are: 47,1% positives; 13.7% negatives; 8,8% false positives, and 30,4% false negatives with an overall correct prediction of 60.8%.

The second predictor is χ_{em}; end mental score of the practical flight selection (β = -0.496, p = 0.063*). The classification results of the model with χ_{em} as predictor are: 46,1% positives; 11,8% negatives; 9,8% false positives, and 32.4% false negatives with an overall correct prediction of 57.8%.

4.5.3 Chance capitalisation

In general classification results can be represented a little brighter than they actually are. This phenomenon is called chance capitalisation. To test for chance capitalisation a backward logistic regression analysis on predictor χ_{em} is performed according to the method of 'leaving one out'. This sample consists of 110 cases; with the leaving one out analysis method a number of 110 analyses could be performed. In the first analysis the 1st case was excluded and 2nd to 110th case included, in the second analysis the second case is exclude but first case and third to 110th case included and so on. Results of the calculations are inserted into the regression model as well as the score of the one case left out. When end result of this model and score is 0.5 or higher the one case left out is placed in the pass group, otherwise the one case left out is placed in the fail group. Missing value cases were excluded leading to 102 cases. If chance capitalisation were to play a role; classification results on overall correct prediction of the leaving one out method will be less than results from the individual analysis. Classification results can be found in Table 5. Classification results indicate that chance capitalization did play a role in the analysis of χ_{em}. The percentages of overall correct classification dropped with 11,3% in the leaving on out method from 57,8% to 46,5%. This result is striking; it seems that prediction without predictors produces a better overall prediction than prediction with a significant predictor.



Classification leaving one out, predictor mental load PFS			Observed	Observed	
			EMFT passed	EMFT failed	
basemodel	**Predicted**	EMFT passed	55,9%	44,1%	100%
pred ment load PFS		EMFT failed	0,0%	0,0%	0%
			56%	44%	
				basemodel overall correct prediction =	55,9
			EMFT passed	EMFT failed	
log regr analysis	**Predicted**	EMFT passed	46,1%	32,4%	78,4%
pred ment load PFS		EMFT failed	9,8%	11,8%	21,6%
			56%	44%	
				log regr analysis overall correct prediction =	57,8%
			EMFT passed	EMFT failed	
leaving one out	**Predicted**	EMFT passed	46,1%	32,4%	78,4%
pred ment load PFS		EMFT failed	9,8%	11,8%	21,6%
			56%	44%	
				leaving one out overall correct prediction =	46,5%

Table 5. Classification results of the base model and results of leaving one out method backward logistic regression analysis on predictor Z_{em}

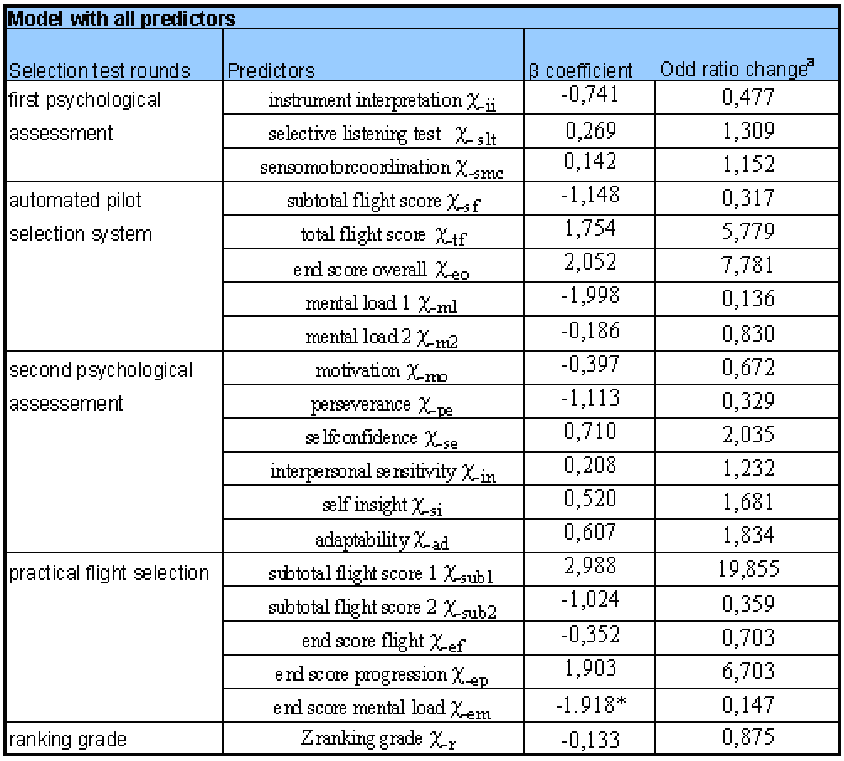
4.5.4 Analysis of a model with all predictors

In reality the RNLAF performs all selection tests and then makes a decision whether a candidate goes on to be a flight trainee or not. To give an idea of what happens when all selection tests are used results are given for a full logistic regression model with all predictors included.

This model shows that when all predictors are included there is one predictor with a significant result; mental load score of the practical flight selection. Its regression weights and odd ratio change can be found in Table 5.

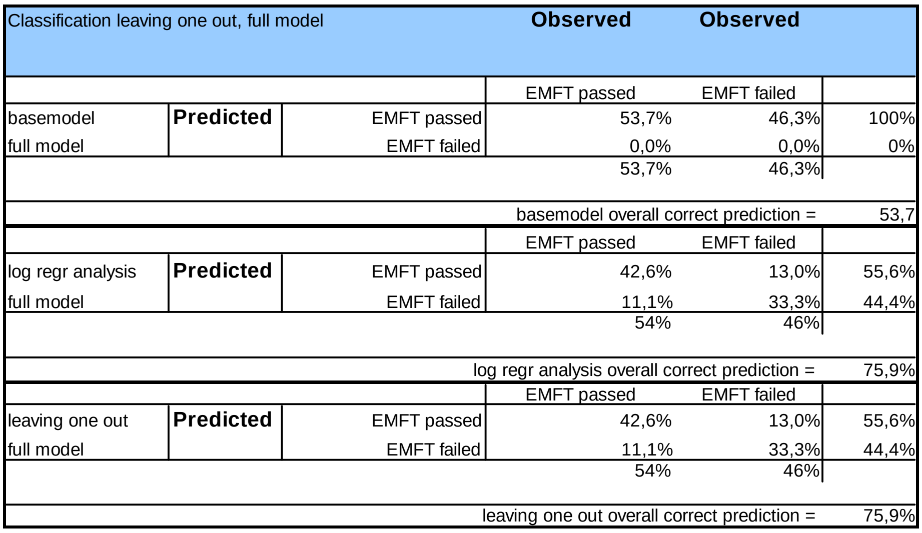
The model with all predictors included has got a good Hosmer and Lemeshow fit $\chi 2$ (8, N = 110) = 4.964, p = 0.761. The pseudo R^2 = 0.434 (Nagelkerke). Pseudo R^2 is around 0.4, indicating that nearly half of the variation is explained by the model.

Classification results indicate that 75,9% of all cases are correctly classified when using the model with all predictors. This model shows 42,6% positives; 33.3% negatives; 13,0% false positives and 11,1% false negatives. Indications for chance capitalization are not found since overall correct classification percentages are alike in regular logistic regression analysis and a leaving one out method. These results can be found in Table 6.



Model with all predictors			
Selection test rounds	Predictors	β coefficient	Odd ratio change[a]
first psychological assessment	instrument interpretation χ_{ii}	-0,741	0,477
	selective listening test χ_{slt}	0,269	1,309
	sensomotorcoordination χ_{smc}	0,142	1,152
automated pilot selection system	subtotal flight score χ_{sf}	-1,148	0,317
	total flight score χ_{tf}	1,754	5,779
	end score overall χ_{eo}	2,052	7,781
	mental load 1 χ_{ml}	-1,998	0,136
	mental load 2 χ_{m2}	-0,186	0,830
second psychological assessement	motivation χ_{mo}	-0,397	0,672
	perseverance χ_{pe}	-1,113	0,329
	selfconfidence χ_{se}	0,710	2,035
	interpersonal sensitivity χ_{in}	0,208	1,232
	self insight χ_{si}	0,520	1,681
	adaptability χ_{ad}	0,607	1,834
practical flight selection	subtotal flight score 1 χ_{sub1}	2,988	19,855
	subtotal flight score 2 χ_{sub2}	-1,024	0,359
	end score flight χ_{ef}	-0,352	0,703
	end score progression χ_{ep}	1,903	6,703
	end score mental load χ_{em}	-1.918*	0,147
ranking grade	Zranking grade χ_{r}	-0,133	0,875

Table 5. Coefficient and odd ratio change of model with all predictors. [a] = the ratio change in the odds of the passing/failing EMFT for a one-unit enhancement of a predictor while all others stay equal. * p < .10



Classification leaving one out, full model			**Observed**	**Observed**	
			EMFT passed	EMFT failed	
basemodel	**Predicted**	EMFT passed	53,7%	46,3%	100%
full model		EMFT failed	0,0%	0,0%	0%
			53,7%	46,3%	
			basemodel overall correct prediction =		53,7
			EMFT passed	EMFT failed	
log regr analysis	**Predicted**	EMFT passed	42,6%	13,0%	55,6%
full model		EMFT failed	11,1%	33,3%	44,4%
			54%	46%	
			log regr analysis overall correct prediction =		75,9%
			EMFT passed	EMFT failed	
leaving one out	**Predicted**	EMFT passed	42,6%	13,0%	55,6%
full model		EMFT failed	11,1%	33,3%	44,4%
			54%	46%	
			leaving one out overall correct prediction =		75,9%

Table 6. Classification results for a model with all predictors, normal logistic regression analysis and results of the leaving one out method

5. Discussion and conclusions

This study was conducted following a research question and hypotheses. It was hypothesized which selection tests have the greatest predictive value on passing/failing the Elementary Military Flight training (EMFT). In addition, hypotheses covered a change in false positives and false negatives and covered added predictive value of predictors.

5.1 Research question

The research question posed by the RNLAF was: *What is the predictive validity of the pilot selection tests of the Royal Netherlands Air Force concerning the chances of succeeding the elementary military flight training for the years 2005 to 2008?*

In total the predictive value of all selection tests is small. A model that includes all predictors contains one statistically significant predictor. The backward analysis showed that the predictor of mental load in the practical flight selection test had statistically significant predictive value in a composite of all selection test scores. In the individual predictor analyses the predictors of mental load and progression (PFS) showed statistically significant predictive value.

5.2 Hypothesis 1: *Using the predictors from the selection tests of the RNLAF causes a change in wrong classification when compared to classification without predictors.*

5.2.1 The first psychological assessment

In all backward logistic regression analyses all but one group (the predictors of the first psychological assessment) of predictors or individual predictors changed the classification results of the model compared to the base model. These results indicate that, when excluding the first psychological assessment, hypothesis 1 can be accepted.

Since the first psychological assessment is the first test round in the selection procedure of the RNLAF an explanation for the results could be that these tests are not predictive of results in the EMFT but are predictive of following selection rounds.

5.3 Hypothesis 2: *The capacities measured in the first psychological assessment are the predictors with the greatest influence on the probability of correctly classifying the pass/fail EMFT criterion.* and 3: *The scores measured in the automated pilot selection system, and the scores measured in the practical flight selection are the predictors with the greatest influence on the probability of correctly classifying the pass/fail EMFT criterion.*

5.3.1 Sample size and sample coincidence

Significant support was not found for hypothesis 2 (first psychological assessment) Hypothesis 2 is rejected.

Furthermore significant support was found for the tests of the PFS but not found for tests from the APSS. Hypothesis 3 is accepted for the PFS tests but rejected for the APSS tests. On the contrary to results for hypothesis 3 a previous RNLAF validity study (RNLAF [3]) and other research (Hunter & Burke, 1995) showed significant results for both APSS and PFS.

A power analysis would be useful. However, constructing a power analysis for a logistic regression model of multiple (twenty) predictors is too complicated and was not performed.

5.3.2 Up or out system in selection procedure

The selection procedure of the RNLAF works via a principle in which scores of a particular selection round decides whether a candidate proceeds into the next round. It might be wiser to look at the predictive validity of selection rounds to its following selection round since that is what selection rounds are decisive on. Significant results in the present study have been found in the selection round of practical flight selection which is closest to the criterion passing/failing the elementary military flight training.

5.3.3 Restriction of range

Restriction of range can have a negative influence on the results. One way of solving this would be to create a control-group of participants in the EMFT that have not passed selection tests of the RNLAF. However, this is not possible.

5.3.4 Comparison with previous RNLAF research

In 2005 the RNLAF conducted a validation study on their selection tests (RNLAF [3]). The results of this study indicated good predictive value for both scores of the automated pilot selection system (APSS) and the practical flight selection (PFS). These results were partly replicated in the present validation study for the PFS scores. In the present study all selection test variables are included in the analysis, whereas in the validation study of 2006 only the results from the APSS and PFS were included. The number of cases was comparable.

Reasons for failing the elementary military flight training were not included in the present study. There was no distinction between trainees that failed due to lack of flight performance and trainees that failed because of other reasons, for example: loss of motivation. Reasons for failing could add information.

5.4 Hypothesis 4: Individual predictors or sets of predictors add predictive value to the base regression model.

5.4.1 Significant predictors

Significant predictors can be found in the practical flight selection. The significant predictors are: χ_{em} and χ_{ep}, leaving 18 other predictors non-significant. Hypothesis 4 can be accepted. Sample coincidence and criterion placement can have its influences.

6. Recommendations

Two sorts of recommendations are distilled. Recommendations for future research will be discussed first. Secondly, practical recommendations will be discussed.

6.1Future research recommendations

6.1.1 Sample size

To gain more knowledge on the predictive value of the selection tests of the RNLAF it is recommended to increase the sample size. One way to extend the sample size is to add cases to the existing sample size with each completion of the elementary military flight training. Annually this would lead to an approximate increase of 30 cases.

The second way to extend the sample size is to add selection score data from selection tests used in other countries and perform a meta-analysis on measured constructs.

6.1.2 Research methods

First, the criterion used in present validation study was passing/failing of the Elementary Military Flight Training (EMFT). It is recommended to conduct a pilot study in which criterion setting on a following test round is tested. This is a method that fits the reality in the selection procedure of the RNLAF perfectly.

Second, in the past the scores from the automated pilot selection system and the scores of the practical flight selection were used for analysis, and scores of the first and second psychological assessment were a priori excluded, whereas in the present study scores from all selection tests were used. It is recommended that scores of all selection tests are analysed.

Third, previous research included reasons why trainees failed the EMFT. A recommendation is to include the reasons for failing the EMFT in future research.

Fourth, in all validation studies of the RNLAF the tests were used in the analyses instead of the constructs those tests measure. It is a given that sample sizes are always small and that norms and testcontent will always change making it difficult to perform longitudinal research or a cross validation. It is recommended to perform a pilot study in which constructs are analysed instead of the selection tests.

6.1.3 Additional research

First, it is recommended to keep the same data set and perform a different statistical analysis on all selection tests to compare those results to present results. Second, it is recommended to plan a repeated logistic regression analysis after each sample extension to detect possible changes in the results.

Possibly several predictors correlate highly; partial correlation. To find out whether this is an issue it is recommended to perform a path analysis.

In general, it seems that a general cognitive factor ‘g’ has a good predictive validity, especially when this general cognitive factor is tested together with other constructs (Tsang & Vidulich, 2008, Hunter & Burke, 1995). It is recommended that the RNLAF performs a pilot study adding measurement of ‘g’ to their selection procedure, or at least consider adding measurement of ‘g’.

6.2 Practical recommendations for the RNLAF

6.2.1 Archiving data

Data was gathered from diverse places and databases. A recommendation to the RNLAF is to construct a central database (digital and/or paper dossiers) where data from all the selection tests, officers training, and elementary military flight training are archived. Paper dossiers are recommended to be archived at a central place with a secured take-out system.

Further, it is recommended to keep an archive of digital data for all selection scores and all performance scores of the elementary military flight training. An advantage of digital data is the ease with which back ups can be made.

This leads to a third recommendation. Back ups of selection scores and performance scores of the elementary military flight training are necessary. It is recommended that periodically a back up of data is planned and performed.

6.2.2 Gathering extra data

During the data gathering of the present study it showed that the elementary military flight school did not archive input and output results of their trainees. It is recommended that the elementary military flight school keeps their own records of input, output and attrition rate.

To improve quality of future validation studies it is recommended to not only keep a record of passing/failing but also to keep a record of flight scores and other performance scores at the elementary military flight school.

6.2.3 Keeping track of selection test changes

An overview of changes that occur in the selection tests of the RNLAF is not readily available. A recommendation is to include information on methodology, test changes, and measured constructs, in the RNLAF psychological selection quality handbooks.

7. References

Burke E., Hobson,C., & Linsky, C. (1997). Large sample validations of three general predictors of pilot training success. *International Journal of Aviation Psychology, 7,* 225-234.

Centrum voor Mens en Luchtvaart [Centre for Man in Aviation]. (2007). *Kwaliteitshandboek Deel II Processen.* (KHB CML II 10-03-08 (1)). Soesterberg, The Netherlands.

Cramer, D. (2003). *Advanced Quantitative Data Analysis.* Open University Press, Philadelphia, United States of America. (119-142).

Defensie beeldbank. (2008). *Een jonge belangstellende in de cockpitsimulator van een F16.* (060629SH3031D) (Retrieved August 1, 2008, from http://defensiebeeldbank.mindef.nl/). Den Haag, The Netherlands: Hilckmann, S.

Delaney, D. (1992). Dichotic Listening and psychomotor task performance as predictor of naval primary flight-training criteria [abstract]. *International Journal of Aviation Psychology, 2,* 107-120.

Dockeray, F.C., & Isaacs, S. (1921). Psychological research in aviation in Italy, France, England, and the American expeditionary forces. *Journal of comparative psychology, 1,* 115-148.

Dunbar, S.B. & Linn, R.L. (1991). Range restriction adjustments in the prediction of military job performance. In A.K. Wigdor & B.F. Green (Eds.), *Performance assessment for the workplace (Volume II).* Washington, DC: National Academy Press.

Giles, D.C. (2004). *Advanced research methods in psychology.* Routledge, London/New York, England/ United States of America. (73-80).

Hunter, D.R., & Burke, E.F. (1995). *Handbook of pilot selection.* Avebury Aviation, Ashgate Publishing Limited, Aldershot, England. (83-128).

Hunter, D.R., & Burke, E.F. (1995) [2]. *Handbook of pilot selection.* Avebury Aviation, Ashgate Publishing Limited, Aldershot, England. (104-127).

Hunter, J.E. & Schmidt, F.L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75,* 334-349.

Koninlijke Luchtmacht [Royal Netherlands Air Force] [1] (2008). Retrieved May 21, 2008, from http://www.werkenbijdeluchtmacht.nl .

Koninlijke Luchtmacht [Royal Netherlands Air Force] [2] (2004). *Information booklet for applicants.*

Koninklijke Luchtmacht [Royal Netherlands Air Force] [3]. (2005).*De kracht van job samples in de selectie van vliegers voor de Koninklijke Luchtmacht.* Centre for Man in Aviation, Soesterberg, The Netherlands: Harsveld, M., Hijmans, R.M.A. & Koning, A.J.

Koninklijke Luchtmacht [Royal Netherlands Air Force] [4]. (unknown) *Instrument Interpretatie Test.* Centre for Man in Aviation, Soesterberg, The Netherlands

Koninklijke Luchtmacht [Royal Netherlands Air Force] [5]. (unknown) *Selectieve Luister Test.* Centre for Man in Aviation, Soesterberg, The Netherlands

Marques, M., & Gomes, A. (2008, May). Prediction of flight success in military pilots: neural networks, discriminant analysis and logistic regression as classificatory methods. Conducted at the XXXII meeting of the Euro-NATO human performance in military aviation working group, Lisbon, Portugal.

Martinussen, M., & Torjussen, T. (1998). Pilot selection in the Norwegian Air Force: a validation and meta-analysis of the test battery [abstract]. *International Journal of Aviation Psychology, 8,* 33-45.

Parker, G., & Oliver, N. (2006). Test review: The Vienna test system. *Journal of Occupational Psychology, Employment and disability. 8,* 169-176.

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

Tactische Luchtmacht [Tactical Airforce]. (2007). *Beleidskader psychologische selectie voor luchtvaartgerelateerde functies voor de Koninklijke luchtmacht.* (Versie december 2007). Soesterberg, The Netherlands: van Nieuwburg, C.M.

The American Heritage (2000). *Dictionary of the English Language: Fourth Edition.* Retrieved February 28, 2008, from http://www.bartleby.com/61/3/T0010300.html

Tsang, P.S., & Vidulich, M.A. (eds). (2008). *Principles and practices of aviation psychology.* Mahwah, NJ, Lawrence Erlbaum. (357-396).

Urdan, T.C. (2005). *Statistics in plain English.* Including CD-ROM. Mahwah, NJ, Lawrence Erlbaum. (145-160).

Appendix A. Regression analyses

A.1 Regression

Regression is used to examine strength and nature of the relations between different variables. It shows predictive power of an independent variable on another dependent variable, or relative predictive power of a set of independent variables on one dependent variable. Regression offers the feature of examining predictive relations of a variable on another variable while controlling for a covariate. Its purpose is to make predictions about an outcome variable based on data of a set of independent variables. Example; when variables on sizes of owned houses and heights of income are known, a prediction on the size of the house might be made when looking at the height of income. Regression analysis produces a formula for calculating the predicted value of one variable when we know the actual value of the second variable. To understand explanations of linear and logistic regression basic knowledge of linear equations, probability calculations, and logarithms is required. When basic knowledge needs to be accessed readers are directed to “Statistics for dummies”[4] for simple and easy accessible explanations on linear equations and probability calculations. Basic information on the workings of logarithms can be found in “Calculus”[5].

A.2 Linear Regression

An assumption made in the world of statistics is that relations between variables are linear. One variable has got the same amount of influence on the other variable. Since statistics are always bare versions of reality, regression is depicted by a model. An example model of single linear regression is visualized in Fig. A1. In the model it can be seen that multiple variables (x) are possible. An example: sizes of houses can be related to height of income. However, it can also

[4] D. Rumsey. (2003). *Statistics for dummies.* Whiley.

[5] J. Stewart. (2007). *Calculus.* Cengage Learning.

Linear regression model:

$$\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_p * x_p$$

b_0 = intercept

$x_{(1,2,\ldots,p)}$ = predictor

$b_{(1,2,\ldots,p)}$ = regression coefficients for predictors

Fig. 1A. A single lineair regression model

be related to the amount of building ground available, and possibly even be related to the culturally defined status of house-size.

When multiple variables are predictors of an outcome variable this is called multiple regression. Multiple regression can show four results: How much predictor variables as a group are related to the outcome variable, the strength of a relationship between each predictor variable and the outcome variable when controlling for the other predictor variables, the relative strength of each predictor variable and lastly, it shows any relations between the predictor variables. In linear regression, or multiple linear regression the outcome variable is measured in quantities (Urdan, T.C., 2005).

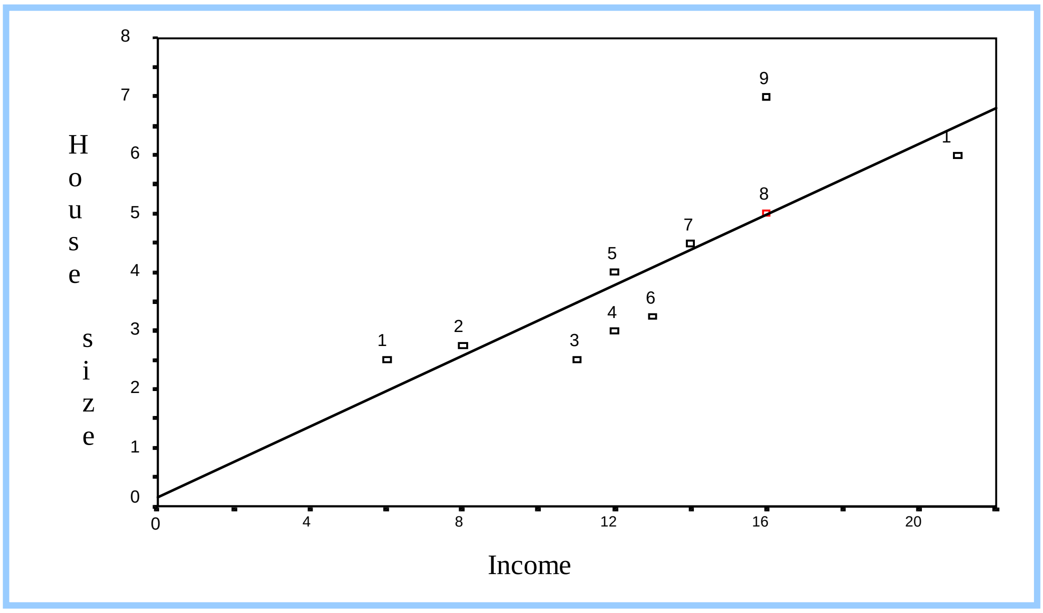


Fig. A1. An example of linear regression, y = income and the predictor variable x = education in years. The dots are scores (1-10). This plot visualizes the relation between the amount of education and the amount of income: when education goes up, income does so to. (CD-ROM, Urdan, T.C., 2005).

A.3 Logistic Regression

Logistic regression is used to make predictions on to which group each case in the study will belong. Based on scores of that case will it belong to one group of a criterion, or belong to the other side (Giles, D.C. 2004)? Example: a house can be either owned or not owned. Predictions are made based on odds or a probability of a case belonging to the own a house-group or do not own a house-group. Probability can vary from a minimum of 0 (no chance at all the house will be owned) to a maximum of 1 (the house is owned for sure). A logistic regression model stands for: the probability of a case belonging to a group (P) is the number of times that case belonging to that group is present divided by the total number of times it could be present. This can be depicted in a model visualized in Figure A2:

Logistic regression model:

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

P = probability of positive result in dichotomous variable
e = base of natural logarithm
a = intercept (compare to b_0 in lineair regression)
X = predictor
b = regression coefficient for predictor

Fig. A2. A logistic regression model

Logistic regression assumes that the relationship between criterion and predictors is best depicted by an S-shaped line, as can be seen in Figure A2, instead of a linear line as in A1.

The relationship in the S-curve is expressed in the log of odds. To rebuild the logs into odds the natural logarithm of e is raised by the power of the log (Cramer, D., 2003).

In short, whereas linear regression uses the regression coefficients and the constant to calculate the predicted value of a case, logistic regression uses regression coefficient and the constant to calculate the odds, expressed in a

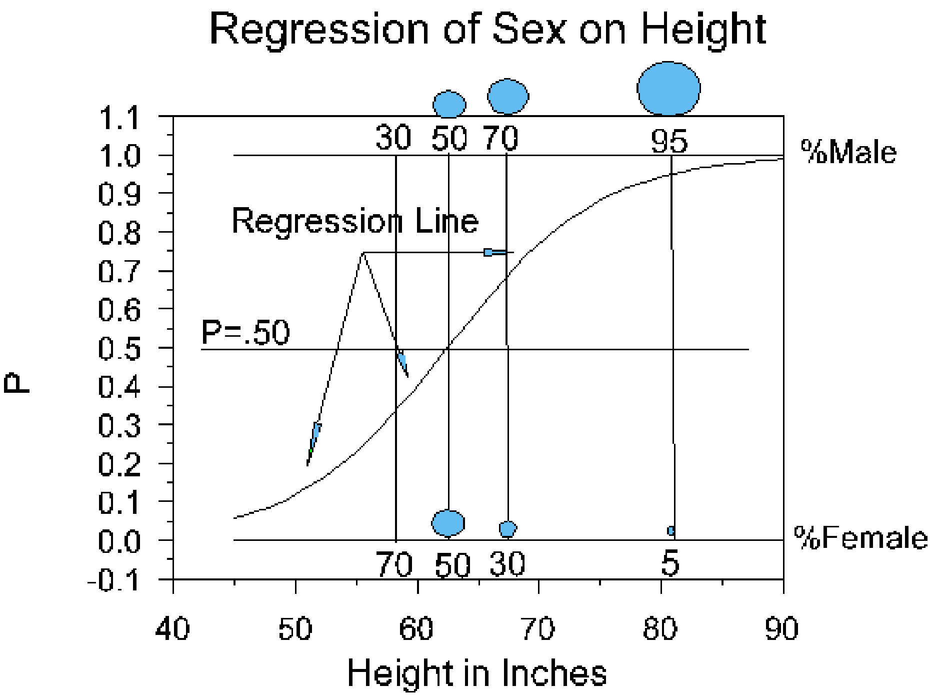


Fig. A2: an example of a graph of logistic regression. Dichotomous criterion$_{\text{gender}}$; either female or male. Height in inches is X. The amount of P is determined by b, a, and e.

logarithm. The logarithm odds are then converted into odds and then odds calculate the predicted probability of a case.

Example: with linear regression it can be predicted what the size of a house is based on the regression coefficient of the income, with logistic regression it can be predicted what the chances are that the house is owned or not based on the regression coefficient of the income.