Selection of offshore rotary wing pilots:

Do psychological tests

predict simulator performance?

A thesis in work and organisational psychology

Psychologist Tore Wingestad

The Norwegian Defence Leadership Institute

Abstract

The purpose of this study was to examine the predictive validity of psychological tests used for the selection of helicopter pilots. Nine psychological tests, cognitive and psychomotor, were used. In addition a psychological interview was conducted. The total sample comprised 100 experienced pilots, of which 67 % passed the evaluation in the simulator, - applied as criterion in the study. Both the un-weighted mean test score and the psychologist's ratings based on the interview predicted pass/fail in the simulator (with respectively (r = .51 and r = .54). The results suggest that psychological tests and evaluations predict simulator performance satisfactorily, and thus that the applied selection procedure is a cost effective method compared to more expensive time in the simulator. It is most probably also a way to predict safety in flight operations.

Preface

For some years now I have been engaged as a consultant in selection of rotary wing pilots for different companies operating offshore. As a final part of my specialist training in work and organisational psychology, I thus decided to explore and do this thesis on the validity of the applied selection procedure. Validity is a complex construct, and in this thesis the assessment and performance in the simulator – which is a final part of the total pilot selection procedure in these companies – is applied as the criterion.

I appreciate this opportunity to express my gratitude to some of those who have supported this study and assisted me in my work. First of all, I have to thank the rotary wing companies operating offshore for their outmost cooperation in all parts and phases. Secondly Dr. Psychol. Monica Martinussen at the University of Tromsoe has been extremely useful guiding me and offering feedback. So have the Armed Forces Chief Psychologist and Research Director, Prof.dr.philos Jon Christian Laberg, and Specialist in work and organisational psychology Tore Torjussen. I will also state my gratitude to psychologist Jon Lars Syversen for his feedback and translation. Without the devoted and skilful practical assistance from Erik Andreassen serving his military duty here at the Armed Forces Institute of Leadership, this thesis would certainly not have been finished by now. Thanks to all of you, and any shortcoming in this final report is of course mine only.

Tore Wingestad
Akershus Castle, February 1., 2005

Introduction

Selection is the construct we apply on the process in which you apply one or a set of methods and test on a sample of persons, aiming to identify those persons which are more likely to be successful in a given education, training, career or position. The decisions regarding which ones to select, are based on the available and relevant information on this given career line or position. It is thus extremely important to verify the real link between the scores achieved in the applied tests/methods and future performances from training or work – namely criterions of a successful selection. This link is what we usually label predictive validity, and it is given in the form of a statistical correlation coefficient ranging from .00 to 1.00, where 1.00 express a perfect positive correlation. The goal of this study is to explore the predictive validity of the methods and procedure applied in the selection of offshore rotary wing pilots, by using a intermediate criterion, namely performances in the rotary wing simulator.

In Norway, as in many other countries, the Air Force is the single institution training and producing the main part of professional pilots, and psychological tests and methods have been used for decenniums. In Norway, such tests and methods were introduced after the 2. World War (Riis, 1955). Since then the test battery has been extended, developed and redesigned a number of times (Torjussen & Hansen, 1999), and also re-validated several times (Martinussen & Torjussen, 1998).

A brief story on the history of pilot selection.

A bit jokingly it is stated that the first pilot selection procedure was applied by Orville and Wilbur Wright in 1903, when they flipped a coin on which one of them to take the first flight with the aircraft they had built. An article in The Lancet from 1918; *"Essenctial Characteristics of Successful Aviators"*, based on observations of pilots during the 1. World War, concludes in the same line and states that pilots were certainly not supermen, but simply good in sports, and topped with good initiative and sense of humour (Turnbull, 1992). The tests applied today are, in spite of still some limitations, clearly more sophisticated, and the result of almost 100 years of systematic trial and error, development and validations studies. (Hunter & Burke, 1995).

To become a pilot has for long been an attractive and popular career choice. The number of eager applicants has been vast, and thus given the opportunity also to select the ones with the highest scores and credentials from schools etc. Still, theses academic scores have proven insufficient. The pilot candidates selected by such measures were clearly not necessarily the most successful ones in pilot training. With this painful lesson, psychologists and psychiatrists in US and Canada started a project and closer cooperation with experienced pilots, aiming to develop more reliable methods with better predictive power. (Storsve, 1983).

The development of the aircraft and the pilot profession has since the 2. World War made most Air Forces apply psychological methods in their selection. The psychological tests and methods have been combined to rather strict medical requirements as well as progression in the basic flight training. The gains by this are clear enough. Firstly, pilot training is expensive and should be given only to those who succeed. Secondly, the consequences of pilot errors are serious, sometimes fatal, and to be avoided as much as possible. The goal was then to identify the kind of skills, aptitudes and personality traits that were important for a successful pilot training and career, and how these possibly could be measured or estimated. Today, the body of research on this is getting solid, an the international community by and large agree on how these measures can be collected and used. The aptitude tests have traditionally been of two categories:

- General abilities: Most commonly paper-and-pencil tests, measuring reasoning, memory, technical-mechanical comprehension, spatial ability etc.
- Psychomotoric abilities: Aiming to measure motoric coordination, reaction time, information processing, simultaneous capacity etc.

Meta-analyses of validation studies have demonstrated good predictive validity for the tests on general and psychomotoric abilities used in pilot selection. (Hunter & Burke, 1995; Martinussen, 1997). A number of tests and questionnaires on personality have been scrutinized in the same way, but with less success; it is far more difficult to demonstrate their predictive validity as a part of pilot selection procedures. (Martinussen, 1997).

Why is selection important?

Why spend time and money on selection of new employees? As in most businesses, the bottom line rules, and it is simply a lot of money potentially saved using a systematic and research based selection method. Equally important, there are a lot of human benefits in ensuring that people are well suited for their position, enjoying and developing their careers. Especially so in the kind of high-risk operations as in the aviation industry, and in selection of pilot and air traffic controllers in particular, as they at times end up as the final filter making critical operative decisions.

The Air Forces all around the world have had greater opportunity to be systematic and to analyse their experiences in selection than most commercial businesses. This due to the great number of candidates tested, checked and assessed for pilot training during especially the last 50 years. Even if the Air Force is a kind of organisation with some unique characteristics also requiring persons with some special assets, a wide range of their experience in selection can be generalised and valuable even for civil purposes. In certain areas, like offshore helicopter operations, the challenges are hardly less than many in the Air Force experience.

The pilot selection procedure applied in the Norwegian Air Force comprises a number of psychological test and formal requirements (Martinussen & Torjussen, 1998; Torjussen & Hansen, 1999). Additionally, a personality test and interview by a psychologist is included. The goal of the interview is to assess motivation and personal suitability, of which communication skills is one important part.

Regarding commercial pilot training, there are no agreed or shared standards, and basically up to the training organisation to set their own standards and requirement for admission. Some flight training schools require only a basic medical check. If then the check rides and examinations are passed, the licences are received, with permission to fly. The airline companies and operators are however free to set the requirements they find necessary. Naturally this implies that there both is and probably will be experienced pilots which actually does not meet these company standards, in spite of having their pilot licences. Further and even more potentially hazardous is the fact that their basic skills and potential for piloting might never be assessed, which common sense suggests as a good starting point – before the professional pilot career is initiated.

The optimal selection procedure, should focus on skills and assets that are more stable, and in fact also basically resistant to development. The goal must be to gain measures on critical areas, which will set the limits for future performance. For example, a pilot candidate with less than required abilities in reasoning or spatial orientation, will basically have this limitation no matter how much training he or she gets.

One might state that it would be most fair if all pilot applicants to the Air force or a commercial airline company, got their chance to prove their skills through a closely supervised training period. However, most of us accept that this would be a less efficient venture, and extremely expensive. Nevertheless, it is a paradox that because pilot training generally is performed in very controlled manners, also less promising assets might be left uncovered, and for different reasons. Hidden limitations during training is however not equal to never appearing in later real flight operations. Applying solid psychological tests and methods in initial selection for pilot training, is therefore an important tool in preventing accidents as well as economical losses for individuals, companies and the society in general.

As a test user or customer, depth knowledge of the actual development of the given test, is not always required. Some basic knowledge of core constructs might work. It is however the responsibility of the test developers to ensure that reliability, validity and norms meet generally acknowledged scientific requirements (Madsen, 1991). At the same time, the test user and even client company do have their responsibilities, for example by interpreting and using the test results according to ethical standards and guidelines. A test might also work well for one sample, while being almost useless for a another.

Reliability and Validity

Reliability is a construct and measure on how accurate and reliable the test or method is working. Does it measure – whatever it measures – accurately, or is the measure achieved more or less by chance? If the same person is measured twice with the same test, other things being equal, the scores should be fairly similar – if the test is accurate. This is one way of checking reliability, and is called test-retest reliability. A classic alternative to this measure, is to split the test in two equal parts, and to compute the correlation between the two ("split – half" reliability) (Hellevik, 1999). There are several other statistical procedures expressing internal consistence.

While reliability refers to accurateness and consistence, validity refers to whether and how well the test is actually measuring what it is intended to measure. This includes the use and consequences, actually including the decisions based on the test result. For example, does the test on IQ measure capacities, or something different, - like number of years attended in the classroom. If it really measures the basic capacity, the intelligence test is said to have "construct validity". As mentioned earlier, another way around this is to work out measures on how well a test is able to predict future performances. This is called predictive validity or criterion related validity, usually attained by comparing test performance with a later reasonable measure on job performance.

Criterions

Selection of personnel concerns giving employment to the ones best or sufficiently qualified for the position. And how do we check if the prognosis is right? As given above, this is achieved by comparing the prognosis or test score with some objective measure related to production, or in some cases even more subjective assessments on performance given by a boss or supervisor. The criterions should however meet some minimum requirements in order to be generally accepted (Madsen, 1999):

- They should be relevant, that is being measures on success or productivity.
- They should address basic aspects, not being overly concerned on details.
- They should be practical in use, available by reasonable means and procedures.
- The criterions should also be measurable in a reliable way and have good construct validity.

The criterions on pilot performance, used to measure validity of pilot selection procedures, have traditionally been pass/fail in pilot training. Only in some very rare cases performance in later phases of a pilot career has been applied as criterion (Martinussen, 1997).

Selection of offshore rotary wing pilots

All candidates tested and assessed for offshore piloting are experienced pilots, because the offshore oil companies as end users (and not the operators themselves) requires a minimum of 1000 flying hours for the pilots transporting their employees. Beside this minimum of flying experience, the candidates traditionally had to pass an interview, a medical check and a simulator session prior to being accepted for the company training as offshore pilots. For some years now, a battery of psychological tests and interview with psychologist has been included in the selection procedure.

The rationale for having rather strict requirements for these offshore pilots, is traditionally a notion of the offshore hardships. The operation is often characterised by poor visibility, rough weather conditions and turbulence, “all white surroundings”, instrument flying conditions and high workload which at times demand a lot of the pilots. Contrary to airplanes approaching and landing more or less straight a head with two pilots in the same loop, the helicopter has left- and right turn landings in which only one of the pilots has the critical visual reference, to mention only some of the differences.

Core issues of this study

The goal of this study is to explore the possible relationships between the psychological methods (tests and interview) applied in the selection procedure, and the performances in the simulator. As given above, the simulator session is a compulsory part of the selection procedure applied by the offshore operators. If the psychological methods can predict simulator performances and pass/fail in that session, then time, money and simulator resources might be saved.

Methods

The sample in this study

The sample comprised 100 experienced pilots, all applicants for two major offshore helicopter operators in the North Sea. The number of female applicants is still very modest, and all applicants where therefore handled as one sample. The applicants were tested in the period 2002-2004. The age varied between 22 to 46 years of age (M = 32.7, SD = 5.7), and the number of total flying hours was between 600 to 9000 hrs (M = 2690, SD = 1970).

Descriptions of the applied psychological test battery

The applied battery comprised 8 different tests. The test scores on every test were converted to the Stanine scale (scores from 1 – 9, where 9 is the highest and best score). The set of norms applied in the conversion from raw test score to Stanine, were based on Air Force data. When computing correlations on test performance, only the raw scores were used in order to expose real variation as much as possible.

Ravens Progressive Matrices, Set II (Raven, 1994): This is a classic test measuring general reasoning ability. The task is to analyse a set of symbols and select one of eight matching the pattern in the given set. The test has 36 tasks with increasing complexity, and is one of the best documented test on predicting success in higher education (Hjerkinn, 1994). This test is widely recognised as a reliable indicator on general intelligence, and is applied by the Armed Forces in selection of pilots as well as to higher military academies.

Consequence of modest scores: Low scorers will more probably have difficulties in managing the required progression in training and learning, as well as handling more complex tasks, and generally be more dependent on high motivation to achieve well.

Series of numbers: In this test the task is to analyse series of numbers and identify the rule or system applied when the given serie was made. Responses are given by filling in the next two following numbers. This test is measuring numerical and logical reasoning.

Consequence of modest scores: Probably more difficulties in handling problem solving, especially new and more complex.

Form Object: Forms are given in numerical order on the left side of a sheet. The task is to identify possible smaller part of the forms, given on the right side of the sheet. The presented smaller parts may be rotated, but not mirrored. This test is measuring mental rotation ability.

Consequence of modest scores: Difficulties in making mental rotations, spatial and perceptual functioning in reading maps, radar etc.

Instrument Interpretation: This test is used to give a measure on the ability to visualise aircraft attitudes, to read and combine information from two instruments. The instruments given are compass and artificial horizon, and the task is to identify which one of five aircraft that is matching the instrument settings. This test is measuring spatial ability and reasoning (Hansen, 1987).

Consequence of modest scores: Probably difficulties in making accurate and swift spatial judgements regarding aircraft position and attitude in three dimensions, and in transforming instrument information into mental pictures of the situation.

Direction Tracker: This is a test giving a measure on the ability to keep track and control on headings like up, down, right and left while under increasing mental workload. The test sheet has 4 columns and 4 rows that give 16 squares, and instructions are given from a CD-player about which square to mark, - with either the left or the right hand. The instructions are given slowly in the beginning, and then increasingly faster. The ability to process information swiftly, and to regain control when missing, are important parts of this test.

Consequence of modest scores: Probably being more prone to make mistakes regarding directions in high workload situations, possibly also to less stress-resistance in general.

Digit Finder: The test sheet has 100 squares with numbers. The task is to identify specific digit-squares as fast as possible. This test is measuring perceptual speed, concentration and short-term memory.

Consequence of modest scores: Needs more time to get the overview and to search for details.

Time Estimation and Spatial Orientation: The two tasks are handled simultaneously. The person tested has to estimate time span of different intervals, while at the same time solve spatial problems. A test sheet is given picturing a number of boxes, the tasks is to count how many times one specific box is touching other boxes.

Consequence of modest scores: These two tests are seen as interrelated, and low scores are generally interpreted as possible limitations in simultaneous capacity; i.e. the ability to read instruments or handle the aircraft, while processing new information and making decision on focus of attention.

Tapping: This task is given during the interview with the psychologist, and it is not a strictly standardized test, but more correctly a method. The candidate has two pencils, and is tracking patterns with both hands. Simultaneously, the candidate has to handle various kinds of mental problem solving. While not standardized as a test, the performance is scored on manual precision and coordination, simultaneous capacity, and stress tolerance.

Consequence of modest scores: To handle the manual operations and mental problem solving at the same time, usually with increasing workload, is a stressing situation in which candidates reacts quite differently. Some are able to keep good control and stay on top of the situation, while others turns unable to solve tasks like “3 times 3”, appears to block mental operations, or the manual operations, - or both. It is partly a question of capacity and of finding useful strategies, but also to regain control after being temporarily lost. All aspects are highly relevant in pilot operations.

A Pilot Prognosis is calculated by the un-weighted mean of all the tests in the test battery, including the three scores from the Tapping session.

Procedure

The data for this study was collected in two batches. The first batch is data from the tests and the interview with psychologist. Then all candidates were accepted for the simulator session, no matter what scores they had from the previous testing. The second batch was collection of the data from simulator performances. The simulator session was arranged soon after psychological tests for all candidates. Experienced rotary wing instructors made the assessments in the simulator, and they had no access to information from the previous tests or assessments made by the psychologist.

The interview & assessment of Personal Suitability

To assess communication skills is widely recognised as crucial part of pilot selection. The interview allows the candidate to expose himself and to state his opinions. The interview also allows the psychologist to explore and probe areas of relevance and interest, sometimes unique for the given candidate. This probing opportunity is obviously greater than any questionnaire can offer. The interview is particularly well suited when exploring two main areas accepted as good predictors for future training and work success: the genuine motivation for the work and the position, and the abilities and skills related to social functioning (Madsen, 1991)

Beside skills in communication, the interview also offers opportunities to observe how the candidate exposes himself in more general terms. Tendencies towards being hesitant, avoiding, distant or insecure when discussing some specific area or in general may be probed and assessed. The candidate may expose his eagerness and energy, or the opposite, and a spectrum of personal assets and characteristics. When the situational conditions are accounted for, the interview allows scores on variables like motivation and outlook, cooperation, self image and insight, communication skills, empathy and possible psychological difficulties.

The Simulator Session

During the simulator session, the instructors assess pilot skills, cooperation, and communication in cockpit. The two companies in this study have somewhat different procedures, but these differences are of little relevance for this study and results.

What the instructors focus on, is overall ability to keep overview and control, whether the candidates are working structured, how they cooperate and manage the workload, if they are able to handle critique, come forward with questions when in doubt etc. Endurance, spirit and potential for fulfilling the role as commander, are also examples of what the instructors are looking for, along with pilot skills and pilot relevant abilities. The candidate operates together with a “co-pilot” and the ability to make efficient use of his co-pilot is a point by itself. Special consideration is given on flying in instrument conditions, where the candidate has no visual reference. This is very often reported as a challenge, due to either shortcomings in their basic flying training, or genuine limitations that may require extended training later on. From their observations in the simulator, the instructors make an overall assessment on cooperation and pilot skills, and conclude with pass or fail.

Results

Out of the total on 100 candidates, 67 passed the simulator session. The results from our study is condensed in three tables. Additionally, two figures are presented, illustrating the relations between stanine person/pilot prognoses and rate of pass/fail from the simulator.

Table 1 gives the means and standard deviations for the tests. Some of the tests, and especially "Instrument Interpretations" were skewed in its distribution of scores, as 47 % of our candidates got one of the two top stanine scores (8 and 9). This was however as expected, considering that the candidates are experienced pilots and very familiar to flight instruments.

Table 1

Descriptive statistical measures on the tests & variables in the battery (*N* = 100)

Tests	*M*	*SD*
Series of Numbers	4.7 (9.7)	2.0 (4.2)
Raven Adv.	4.6 (22.9)	2.2 (6.0)
Form Object	5.7 (23.3)	1.9 (7.9)
Instrument Interpretation	6.8 (45.6)	2.0 (12.5)
Direction Tracker	5.3 (35.4)	2.2 (15.4)
Digit Finder	4.6 (37.9)	2.0 (10.3)
Spatial Orientation	6.4 (13.8)	2.2 (9.2)
Time Estimation	4.6 (58.5)	2.1 (33.6)
Tapping (Rhythm/Motoric)	5.6	1.9
Tapping (Simul. Capacity.)	5.3	2.1
Tapping (Stress tolerance)	5.3	2.1
Sum Tapping	5.4	2.0
Personal Suitab. (Interview)	5.7	2.1
Pilot Prognosis	5.3	1.4

Note: Raw scores (in brackets) are transformed to the stanine scale (1-9).

Table 2

Intercorrelations between predictors (N = 100)

Tester	1	2	3	4	5	6	7	8	9	10	11	12	13
1 Series of Numbers	-												
2 Raven Adv.	.64**	- -											
3 Form Object	.23*	.38**	-										
4 Instrument Interprt.	.37**	.44**	-.13	-									
5 Direction Tracker	.55**	.55**	.19	.54**	-								
6 Digit Finder	.34**	.47**	.29**	.32**	.47**	-							
7 Spatial Orientation	.40**	.51**	.25*	.27**	.42**	.20*	-						
8 Time Estimation	-.25*	-.18	-.07	.02	-.11	-.09	-.06	-					
9 Sum Tapping	.51**	.59**	.42**	.33**	.61**	.38**	.43**	-.11	-				
10 Personal Suitabil.	.51**	.58**	.33**	.42**	.56**	.42**	.38**	-.13	.75**	-			
11 Pilot Prognosis	.73**	.82**	.44**	.55**	.79**	.62**	.63**	-.29**	.77**	.73**	-		
12 Flying hrs (n=54)	-.09	-.16	-.32*	-.05	-.25	-.27*	-.22	-.09	-.28*	-.39**	-.29*	-	
13 Age	-.20*	-.25*	-.14	-.17	-.32**	-.32**	-.05	.06	-.29**	-.31**	-.33**	.71**	-

Note**:** *p* < .05. ***p* < .01 (two-tailed).

Table 2 gives the intercorrelations between the predictors. The assessment of performances in the Tapping situation is originally split in the earlier mentioned three variables (rhythm/manual, simultaneous capacity and stress tolerance). These three variables were however highly intercorrelated (.90), and thus computed into one variable only. The different aptitude tests were as expected also positively intercorrelated, except from Spatial orientation and Time Estimation. These two tests are administered as one test and simultaneously; the candidates estimates time span while solving spatial problems. Candidates tend to focus on one of these tasks, and thus the statistical power and stringency of these measures may be more easily obscured. The negative correlation between flying experience (number of flying hours) and most of the predictors was somewhat surprising. Candidates with more flying experience is in fact generally less successful in their test performances.

Table 3

Predictive validity of tests and psychologist assessments, given by correlation to simulator performances (criterion)

Tests	**Simulator-performance (*r*)**
Series of Numbers	.41**
Raven Adv.	.44**
Form Object	.20*
Instrument Interpretation	.33**
Direction Tracker	.38**
Digit Finder	.26**
Spatial Orientation	.27**
Time Estimation	.02
Rhythm/Motoric skills	.56**
Simultaneous Capacity	.55**
Stress Tolerance	.57**
Sum Tapping	.59**
Personal Suitability	.54**
Pilot Prognosis	.51**
Flying hours (n=54)	-.23
Age	-.18

Note. **p* < .05. ***p* < .01 (two-tailed).

All tests except Time Estimation were statistically significantly correlated to the criterion. Raven Adv and Series of Numbers with respectively .44 and .41, and the computed sum Tapping score in particular (.59). The overall assessment from the interview correlates almost equally high (.54). Age and flying experience, as measured by flying hours, does not seem to predict simulator performance. There is in fact a slight negative correlation between these predictors and the criterion (-.18 and -.23, respectively).

The relation between Pilot Prognoses and portion of candidates who passed the simulator session, is given in Figure 1, by percent shares of candidates who passed for each of the stanine score levels. Figure 2 illustrates the same relation for the stanine scores given on Personal Suitability. These illustrations give a rather clear picture of the potential benefit of using these psychological tools in the selection process.

Figure 1

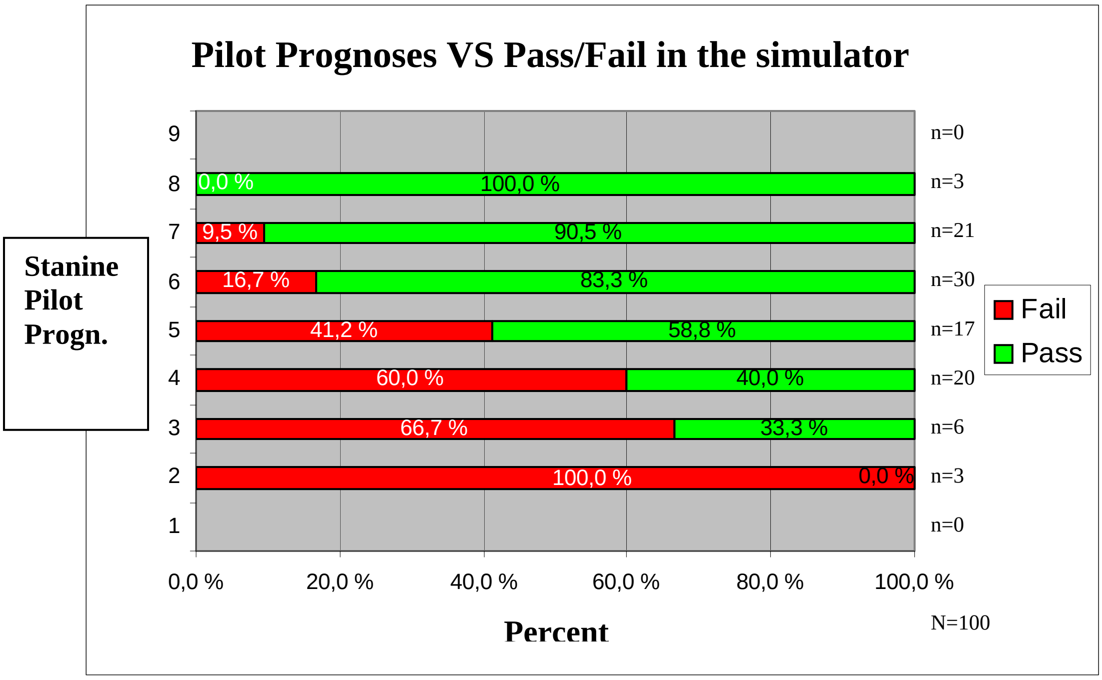

Figure 1 shows that no candidate with lower Pilot Prognose than stanine 3 did pass the simulator session, and that share of candidates is increasing up to stanine 8, where all candidates pass.

Figure 2

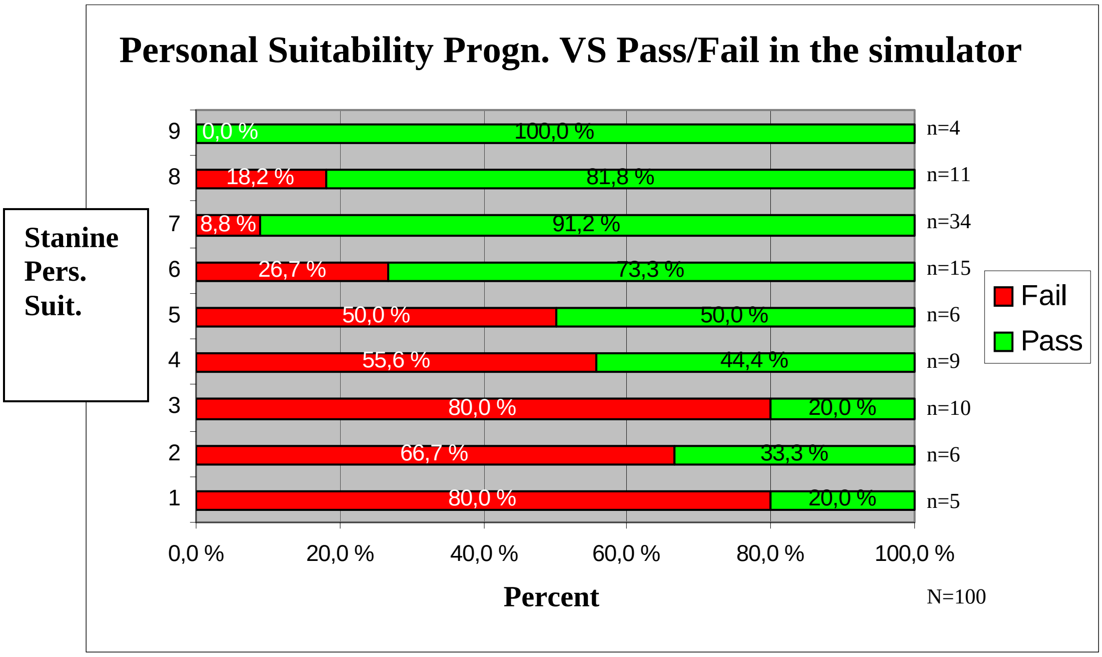

Figure 2 shows that all candidates with stanine 9 as Personal Suitability score passed the simulator session. The percent share passed is reduced as the Personal Suitability scores get lower. Taken together, 2/3 of the candidates with stanine scores 1-5 did not pass the simulator session.

Discussion

In this study, we find that the applied psychological tests and methods are predicting rather well the chances of handling the tasks in a simulator session when applying for the offshore rotary wing companies. Even if a larger part of the candidates do pass in the simulator, a significant portion fail (33 %). All those who failed had commercial pilot training. The sample in this study also included some with Air Force training on rotary wing. This portion was too small to be handled as a subsample for statistical analyses on its own, but none of the candidates with military background failed in the simulator. These candidates are after all strongly preselected on psychological tests when accepted to the Air Force, and should have good potentials to handle also our procedure and the tasks in is simulator session.

The goal of this study was to explore the relations between scores on tests/personal suitability compared to simulator performance. With a reasonable or high correlation, the total selection procedure could be made more cost effective, by excluding a share of the candidates with low stanine scores from the simulator session. The results indicate a rather high correlation between predictors (tests, pilot prognoses, personal suitability) and the criterion (the simulator session).

What then, about the pilots who did not perform well enough in tests and interviews, but handled the simulator well enough? As mentioned, all those included in this study are experienced pilots. This implies that they can be expected to be more familiar with the simulator setting, than by the table with paper-and-pencil test administered by a psychologist. Fair enough, but the psychological tests do measure more specific aptitudes than what can be exposed in a simulator session. The tests are developed to measure stable basic capacities. It would be reasonable to argue that pilots below a certain standard on relevant capacities, more easily would have difficulties in handling demanding operative challenges. From this point of view, low-performers should hardly be accepted for the company, in spite of acceptable performance in a simulator session.

Recruitment based on pilot licenses and number of flying hours only, is not sufficient in this context. The risks by accepting pilots, who are not apt to handle the responsibility, are too high in terms of outcome and costs. In the described selection procedure, some of the candidates have modest stress tolerance, capacity for handling simultaneous tasks, or spatial abilities. At times, we did observe almost collapses in cognitive and psychomotor functioning, and it is very hard to accept that these observations have no relevance for the ability to perform in possible demanding situations in the future. Some of the candidates also did expose modest self insight, emotional capacities and potential for the kind of cooperation required in a small office like a cockpit. If verified and valid, such limitations should also be excluding regarding a future demanding pilot position, and in these cases neither good capacities nor acceptable single simulator performances can compensate. Intellectual capacities can by itself not make up for missing relational abilities.

Some of the explanation behind modest test scores and simulator performance, is probably the fact that many of these candidates have not been to any psychological testing before they started their pilot training. The requirements all the way from being accepted at flying training school to getting the pilot licences, are by and large based on medical and not psychological knowledge. This is contradicting well established international research, experience and consensus in which personal suitability, capacities and psychomotor skills are widely accepted as important factors in order to optimise the chances of success in demanding situations (Fallucco, 2002).

From Table 3, we find that it is the sum of the test procedure that has the best predictive validity. The single tool that has the highest correlation to passing simulator session, is Tapping. As stated earlier, Tapping can not be characterised as a psychological test, as it not standardised and validated according to scientific requirements and guidelines. Tapping is still an important and esteemed tool in pilot selection, both in the Norwegian, Danish and Swedish Air Force. Its relatively high predictive power might be explained by the fact that it resembles a cockpit-like situation more than other tests. The candidate is no longer an anonymous participant in a classroom working with pencil and paper. On the contrary; he is under direct observation and very much aware of being assessed. Furthermore the workload is getting high, problems have to be solved, new instructions are coming, and the attention has to be on the manual task as well. This obviously requires a set of capacities, and also abilities regarding organising oneself and staying on top of the situation.

Earlier meta-analyses of validation studies have given that the mean correlation between predictor-criterion, is .22 for cognitive tests, .20 for psychomotor tests, and more modestly .13 for measures on personality (Martinussen, 1997; Hunter & Burke, 1995). A very likely reason for achieving higher correlation coefficients in this study, is that the sample is stable through the process from collecting test/predictor data and through the final collection of criterion data. In most studies, and alike the procedure in the Norwegian Air Force, only the top candidates in the initial testing and selection process, are allowed to pass for the test period at the flying training school, - were criterion data usually is collected.

This strong reduction of the sample, gives less variance when criterion data is collected. This restriction of range again gives that the observed correlation between tests and criterions, by mere statistical reasons, will be reduced. This is not the case in our study, where all tested candidates went to the simulator session.

Our study furthermore confirms that general cognitive abilities as measured by tests like Raven and Series of Numbers are predicting simulator performance better than psychomotor tests, with Tapping as the exception. This is in accordance to findings in a number of studies evaluating predictors for success in education and career (Madsen, 1991).

In earlier studies, flying experience has also often turned out to be a good predictor on pilot performances. It might be somewhat surprising then, that in our study we find that total flying hours is not correlated to simulator performance. One hypothesis is that “older”, experienced pilots simply are working slower on the tests, while the age and experience is more of an asset when entering a familiar simulator setting. We find however that flying experience correlates little with both the tests and the simulator performance. When experience does not influence simulator performance, it might be caused by variable standards in the often smaller companies and operators the candidates have been employed by earlier in their pilot careers. This is however not necessarily the case, as earlier meta-analyses mainly refers to studies on ab-initio pilot selection. In that context only few of the candidates have some experience. It seems reasonable to expect that flying experience in this context might serve as a positive predictor, while experience actually is of less importance and not equally well predicting pilot performance when the sample is characterised by having usually more than 1000 flying hours. And especially so, if the companies this experience is accumulated in, have less resources for training and investments in maintaining optimal safety-oriented procedures and working cultures.

The variation in how the candidates expose, present themselves and communicates is significant. Some are coming forward with good confidence and eagerness, balanced and well functioning in all respects. Those who are assessed as less suitable based on the interview, generally have basic difficulties in sharing their assets, experiences and ways of functioning with the interviewers. Some have difficulties in communication, appears distant and literally have to be pulled through the interview. In some cases they are not convincing in terms of realistic insight in own functioning and how they relate to others, in other cases we observe behaviour which appears close to depressive tendencies.

When the candidate has to spend extra energy on managing psychological or social-relational tensions or difficulties, mental resources are tied up. Even if there will be situational factors that support or reduce the amount of energy spent on this, we usually infer that there is a high risk of having less resources available for demanding situations in the cockpit and work. Limitations in communication and cooperation is a well known critical factor in the cockpit, and is therefore one of the focus areas in the interview.

The interview is often criticised as a scientific method in assessments. We know however that the validity is improved when the interview is well structured, and when the opportunity of follow-up questions is used. Good questions simply aids the candidate in coming up with relevant information (Hermans og Mulder, 1998). Experience in interviewing is critical, and there are a number of potential traps, like stressing the value of the first impressions and hand shake, applying easy-to-use stereotypes, appreciating some traits and interests too much because they happen to be shared by the interviewer etc. (Hunter & Burke, 1995). Such effects may naturally influence the value and predictive power of the interview a lot.

In this study, the interview-based assessment on Personal Suitability predicted pass/fail in the simulator quite well (.54), and at about the same level as Tapping (.59). Communication and personal assets are however also exposed and evaluated in the simulator session, and should support a positive correlation.

Summary

The goal of this study was to explore how the psychological tests and selection procedure predicted pilot performance. A sample of 100 applicants for offshore rotary wing operation were tested, and the results demonstrate a good correspondence between test measures and the performance in the simulator. The sample in this study is limited, but still sufficient for reliable statistical analyses, and results are clear enough. The question is what implications they might have. The most important implication is perhaps that pilot training and experience is no guarantee for having the qualifications required by major operators. These qualifications are not necessarily developed by operating an aircraft per ce, which might be a useful reminder to operators recruiting pilots.

To be able to predict simulator performance is a good start, but the overall goal is naturally to predict future and long term operative performance. A natural next step in practical research is to study the relation between the selection procedure and advanced company training, operative periodic checks, evaluations on pilot functioning, and career developments markers like succeeding in becoming an aircraft commander.

We also recommend a structured approach to explore simulator performances. Beyond pass/fail, a detailed report on the different variables and assessment areas would be very useful. Especially when the candidate fails, we could learn more by knowing whether this is due to limitations in communication skills, high workload management, basic pilot operative performances or some other area.

Another conclusion is to recommend those who aspire for a pilot career, to have their capacities, psychomotor skills and personal functioning assessed before they enter an expensive and specialised education like pilot training. By such an initial assessment of pilot potentials they would gain personally in several ways. It would also gain training institutions and future employing operators, and most importantly it would increase flight safety in general. At the present, the aviation authorities requires only a medical check prior to pilot training, but a reliable psychological examination is most probably equally important. When not required by the authorities, this end up as a responsibility for the operating companies or individuals, which probably is a less favourable solution. By missing a stringent pre-selection procedure, the production of pilots is much of a “flip a coin” venture, like it was in the days of the Wright brothers, and it should of course be avoided.

Finally we stress again that when the pilot has succeeded in getting operative experience, the employing companies should be careful in using the number of flying hours as a selection criterion by itself. Flying hours is only an asset when other requirements are met.

List of references:

Fallucco, S. J. (2002). *Aircraft Command Techniques*. Aldershot: Ashgate.

Hansen, I. (1987). Beskrivelse av testene. *Militærpsykologiske Meddelelser (MPM)*, 15. Oslo: Forsvarets Psykologiske og Pedagogiske Senter.

Hellevik, O. (1999). *Forskningsmetode i sosiologi og statsvitenskap*. Oslo: Universitetsforlaget.

Hermans, P. H. & Mulder, H. W. (1998). Job analysis and the selection interview. Editor K. M. Goeters, *Aviation Psychology: A Science and a Profession*, (pp.81-92). Aldershot: Ashgate.

Hjerkinn, O. (1994). Det er forskjell på folk. *LUFTLED Norsk Luftmilitært Tidsskrift*, mars, 25-28.

Hunter, D. R. & Burke E. (1995). *Handbook of Pilot Selection*. Aldershot: Ashgate.

Madsen, J. P. (1991). *Systematisk personaludvælgelse: Teori og praksis*. Rungsted: Forsikringshøjskolens Forlag.

Martinussen, M. (1997). *Pilot Selection: A validation and meta-analysis of tests used for predicting pilot performance.* Doktoravhandling. Universitetet i Tromsø.

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

Raven, J. (1994). *Occupational Users Guide: Raven's Advanced Progressive Matrices and Mill Hill Vocabulary Scale*. Oxford: Oxford Psychologists Press.

Riis, E. (1955). Psykologisk utvelging av flygere. *Militærpsykologiske Meddelelser (MPM*), F-2. Oslo: Forsvarets Psykologiske og Pedagogiske Senter.

Storsve, O. (1983). *Seleksjon av flygere til Luftforsvaret i Norge.* Magisteravhandling. Universitetet i Oslo.

Torjussen,T. & Hansen, I. (1999). Forsvaret best i test? Bruk av psykologiske tester i Forsvaret, med spesiell vekt på flygerseleksjon. *Tidsskrift for Norsk Psykologforening,* 8, 772-779.

Turnbull, G. J. (1992). A Review of Military Pilot Selection. *Aviation, Space, and Environmental Medicine*, 63, 825-830.