THE VALIDATION OF THE SELECTION BATTERY FOR PILOTS OF THE SOUTH AFRICAN AIR FORCE

Francois Servaas De Kock

Assignment presented in partial fulfilment of the requirements for the degree of Master of Commerce at the University of Stellenbosch.

Study Leader: Mr Anton Schlechter

April 2004

Declaration

I, the undersigned, hereby declare that the work contained in this assignment is my own original work and that I have not previously, in its entirety or in part, submitted it at any university for a degree.

Signature:

Date: 5 March 2004

ABSTRACT

The recent procurement of modern fourth-generation fighter aircraft by the South African Air Force (SAAF), severe budget constraints, as well as demographic transformation of the South African National Defence Force (SANDF) impacted heavily on the selection and training of SAAF pilots. Against this backdrop, this predictive criterion-related validation study attempted to find an optimal battery to predict various aspects of pilot training performance, using all SAAF qualified pilots from 1997 to 2002 as the sample (N=107). Multiple regression analyses were performed to construct a model which can be used to predict the success of trainee pilots in three phases of pilot training, namely officers' formative training, ground school training and practical flight training. Stepwise regression analyses with training grade achieved as criterion were performed on the data for each of the phases of training. Multiple correlations of 0,34 ($p<0,001$), 0,21 ($p>0,05$) and 0,22 ($p<0,05$) were obtained for flight, ground school and formative training results respectively. Various recommendations regarding the present composition of the battery are made.

OPSOMMING

Die onlangse aanskaffing van moderne vierde-generasie vegvliegtuie deur die Suid Afrikaanse Lugmag (SALM), sowel as omvattende begrotingsbeperkinge en die demografiese transformasie van die Suid-Afrikaanse Nasionale Weermag (SANW) het 'n swaar impak op die keuring en opleiding van SALM vlieëniers gehad. Teen hierdie agtergrond het hierdie voorspellende kriteriumgerigte valideringsstudie gepoog om 'n battery saam te stel wat die verskeie aspekte van prestasie tydens vlieëniersopleiding optimaal kon voorspel. Al die SALM vlieëniers wat gekwalifiseer het van 1997 tot 2002 is in die steekproef ingesluit (N=107). Meervoudige regressie-ontledings is uitgevoer om 'n model te bou wat die sukses van kandidaatvlieëniers kon voorspel tydens die drie fases van opleiding, naamlik offisiersvorming, grondskool en praktiese vliegopleiding. Stapsgewyse regressie-ontleding is gedoen vir elke fase van opleiding, met opleidingspunt behaal as kriterium in elke fase. Meervoudige korrelasies van 0,34 ($p<0,001$), 0,21 ($p>0,05$) en 0,22 ($p<0,05$) is verkry vir vlieg-, grondskool-, en vormingsopleidingspunt onderskeidelik. Verskeie aanbevelings in verband met die samestelling van die battery word gemaak.

Table of Contents

Acknowledgments

Hereby I would like to thank the following for their support in the completion of this thesis:

My Lord Jesus Christ for giving me the opportunity and ability to come this far.

My parents, for motivating me to study.

My study leader, Anton, for his guidance, support, approachability and flexibility.

INTRODUCTION

Military pilot selection has traditionally generated vast amounts of research (Hunter & Burke, 1994). This can be attributed to various factors. Pilots play a key role in modern warfare, and immense costs are involved in their training in terms of both finances and time. In the United Kingdom, the estimated unit cost of training a fast jet pilot is in excess of £3,7 million. In South Africa, the duration of training for one fighter pilot in the South African Air Force (SAAF) takes at least 5 years. Moreover, training failures are costly, where dropout rates are high in the United States Air Force (20%) and Australian and Canadian programmes (30%)(Bourn, 2000). Finally, the costs of aircraft accidents can be considerable in human, financial and psychological terms. For these reasons the military forces conduct ongoing studies to identify effective selection measures.

In the specific context of the SAAF, these concerns are exacerbated by severe budget constraints and the fact that it is currently revamping its aircraft fleet with modern fourth-generation aircraft, with this in itself having its own concomitant implications for human resource selection and development.

In the macro environment various arguments accentuate the need to establish the validity of selection procedures, justifying validation from a pragmatic, scientific and legal perspective.

The pragmatic argument emanates from the fact that organisations are increasingly learning that human resources, where the individual and his/her output is key, are critical to success. Gatewood and Feild (1998) state that "the performance of employees is a major determinant of how successful an organization is in reaching its strategic goals and developing a competitive advantage of rival firms" (p. 3). Selecting people that are likely to perform effectively is a key responsibility of the human resource function, which by implication includes developing and validating effective selection procedures (Campbell, McCloy, Oppler & Sager, 1993; Milkovich & Boudreau, 1997).

From a scientific perspective, it is critical that the selection process is reliable and that it makes valid claims. According to internationally accepted principles and guidelines (American Psychological Association, 2003; United States Department

of Labor, 1978) a sound selection procedure is one that allows valid inferences to be made regarding future job behaviour from available measure scores. Likewise, the Guidelines for the Validation and Use of Assessment Procedures for the Workplace (Society for Industrial Psychology, 1998) concur by stating that the evaluation of any assessment procedure should be "based on the fact that sufficient proof can be found that the procedures used are indeed relevant to the position or work concerned" (p. 1).

The "proof" referred to above can be termed validity. Validity refers to the "degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (American Educational Research Association, American Psychological Association & National Council for Measurement in Education, 1999, p. 184). Validation, therefore, involves the accumulation of evidence - content, criterion or construct-related - to provide a sound scientific basis for the proposed score interpretations (APA, 2003).

From a legal perspective validation is required by law in South Africa, as stipulated in the Employment Equity Act (R.S.A., 1998):

> Psychological testing and other similar assessments of any employee are prohibited unless the test or assessment being used has been scientifically shown to be valid, reliable; can be applied fairly to all employees; and is not biased against any employee or group. (p. 10)

This legislation aims to ensure that the integrity of selection procedures is investigated, especially where certain demographic groups are at risk of being disadvantaged by their use. Furthermore, increasing criticism and pessimism exist about psychometric assessment from the side of labour unions and government (Cook, 1999). The primary concern in this regard relates to the fairness of the selection procedure. Where bias in the test or in prediction is present, the fairness of the use of the procedure may be compromised (Arvey & Sackett, 1993).

In summary, proper validation and use of selection procedures is essential for pragmatic, scientific and statutory reasons. It is appropriate at this point to elaborate on the assumptions underlying the validation process, within the context of personnel selection.

Gatewood and Feild (1998) define selection as the "process of collecting and evaluating information about an individual in order to extend an offer of employment" (pp. 3-4). According to them, the primary function of selection is to separate - from a pool of applicants - those with the appropriate knowledge, skills and abilities (KSAs) to perform well on the job.

During the selection process a choice is made about desired qualities and traits. This choice rests upon a predictive hypothesis that is formulated after considering the demands and context of the job (Guion, 1965). The focus of selection research is then to "test" the predictive hypotheses that certain qualities and traits predict certain desirable behaviour. In this sense, validation is seen as a process of traditional hypothesis testing (Binning & Barrett, 1989; Landy, 1986).

Traditionally, validation research has received considerable attention in the military (Cook, 1999; Rumsey, Walker & Harris, 1994; Schmitt & Borman, 1993). A survey of contemporary literature (Hilton & Dolgin, 1991; Hunter & Burke, 1994) reveals a thorough understanding of the task-demands-KSAs link for the job of the military pilot.

There is general consensus that the determinants of pilot success resort in three main domains, namely intelligence and aptitude, psychomotor coordination, and personality (Carretta & Ree, 1989).

Intelligence and aptitude

Hilton and Dolgin (1991) remark that "there is little doubt that above average intelligence is necessary to master military pilot training"(p. 94). They also characterise intelligence as the best and most stable predictor of flight training success in their summary of pilot selection research during the last century.

Intelligence is a broad concept, and is sometimes defined more specifically. For instance, Ree and Carretta (1996) make a useful distinction between two types of intelligence. They use Spearman's (1904) two-factor theory of cognitive ability and argue that intelligence can be seen as general cognitive ability (*g*) on the one hand, or in terms of specific abilities (s_n) on the other. The factor *g* is a general factor that is obtained though factor analysis and is thought to underlie most of the other intellectual abilities (Plug, Meyer, Louw & Gouws, 1988).

The predictive validity of these types of intelligence appears to differ. Hunter and Burke (1994) - in their meta-analysis - found that general intelligence was not generalisable across studies as a predictor; at most it had an influence moderated by other variables. However, general cognitive ability has consistently been shown to predict pilot training success, showing average statistically significant correlations of 0,33 (Ree & Carretta, 1996).

General intelligence in other forms has also been shown to predict pilot training success. In this line of thinking, it could be argued that *g* is congruent to the construct of fluid intelligence of Cattell (Raven & Court, 1998). Fluid intelligence is defined as intellectual abilities that are determined primarily by genetic factors (as opposed to cultural or environmental factors)(Plug et al., 1988). It could therefore also be expected to predict pilot training success. In this regard, some evidence has been found on the ability of information processing capability - an important indicator of fluid intelligence - to predict pilot training success (Damos, 1996). A more recent South African study has also confirmed this finding, where it was found that pilots could be differentiated from non-pilots on the grounds of rate of information processing (Barkhuizen, Schepers & Coetzee, 2002).

With regard to specific intelligence (s_n), a multitude of abilities have been found to predict pilot training success, amongst others verbal, quantitative, spatial, and mathematical ability, as well as perceptual speed and instrument comprehension (Burke, Hobson & Linsky, 1997; Carretta & Ree, 1996).

The relative importance of *g* and s_n in predicting pilot training success remains a controversial issue. On the one hand, some authors (Burke et al, 1997; Carretta & Ree, 1996) maintain that *g* remains a better predictor of pilot success than specific abilities. Other researchers (Hunter & Burke, 1994; Martinussen, 1996) come to different conclusions and report – as a result of their meta-analyses – that measures of general intelligence had low mean validities compared to more specific measures of intelligence.

Carretta and Ree (1996) add to the debate by stating that the inclusion of specific abilities (s_n) adds little to the ability to predict criteria (see also Ree & Carretta, 1996). Their explanation for their point of view was that many of the additional measures that are used are saturated with *g* and do not represent unique abilities.

However, some authors, for instance Martinussen (1996), are of the opposing view and show that the inclusion of specific abilities indeed had incremental validity over and above measures of *g*. Clearly, the debate on the role of intelligence and aptitude in the prediction of pilot training success is still very active and can be interpreted as an attestation of its dominance in pilot selection batteries.

Psychomotor coordination

Psychomotor skills research has a long history in pilot selection (Griffin & Koonce, 1996). The term "psychomotor" denotes a combination of physical and psychological activities (Plug et al., 1988). Measures of psychomotor coordination - or hand-eye coordination as it is sometimes to referred to - are commonly included in selection batteries for two apparent reasons, being (a) their obvious relation to the task, and (b) the results of validation research that support their inclusion in selection batteries.

In their study, Burke et al. (1997) found that psychomotor tests were predictive of pilot training success and that its validity generalized across samples. They used validity generalization analysis (VGA) with three samples from different national air forces, with a total combined N=1760. A continuation of these authors' findings is the fact that various studies report that measures of psychomotor abilities were able to increase predictive validity of a battery already measuring *g* (Ree & Carretta, 1996). For instance, in one study when psychomotor tasks were added to a USAF selection battery already including the Air Force Officer Qualifying Test (AFOQT) scores, the predictive validity of the battery increased from 0,168 to 0,207 (Damos, 1996).

New developments in psychomotor predictors also abound. For instance, various studies have illustrated the role of situational awareness in pilot functioning (Carretta, Perry & Ree, 1996). Therefore, it can be expected that this construct might prove useful in future pilot selection batteries.

Personality

Personality can be defined as those aspects of individuals that make predictions about their behaviour in specific situations possible (Plug et al., 1988). Contrary to expectation, most studies report that personality adds little to the prediction of pilot success (Carretta & Ree, 1989; Hunter & Burke, 1994; Retzlaff & Gibertini, 1987;

Turnbull, 1992). However, some studies did in fact report that certain aspects of personality had incremental predictive validity in traditional batteries, for instance attitude to risk (Ree & Carretta, 1996). In another study, Carretta (2000) found that a measure of conscientiousness incremented the multiple correlation coefficient of a battery measuring general mental ability from 0,51 to 0,60.

Despite the generally weak ability of personality to predict pilot training success, it is often used in pilot selection. For instance, certain militaries use personality as a screening variable to identify clinical dysfunction and other undesirable traits. It also appears that personality is receiving increased attention in the important areas of stress tolerance and motivation (Hilton & Dolgin, 1991).

Other findings related to the use of personality in selection for training is that, in one study that compared the personality profiles of pilots to those of college students through cluster analysis, it was found that pilots had distinct personalities that distinguished them from non-pilots (Retzlaff & Gibertini, 1987). A similar finding was obtained by a study comparing the personality profiles of student naval pilots to normative data (Lambirth, Dolgin, Rentmeister-Bryant, & Moore, 2003). Ashman and Telfer (1983) found pilots to be more achievement oriented, outgoing, active, competitive, dominant and less introspective, emotional, sensitive and self-effacing than a sample of non-pilots.

In another study, pilot trainees completed a personality inventory measuring five dimensions thought to be associated with flight training performance. After their training was completed, three of the measures were in fact related significantly to training outcome, namely hostility, self-confidence, and values flexibility. Disappointingly, incremental validity analysis did not indicate that the inventory could enhance a selection model already containing traditional aptitude scores (Siem, 1992).

Meta-analyses of predictors of pilot training performance

Hunter and Burke's (1994) meta-analysis of sixty eight published studies, with a total of 437 258 combined cases using the method proposed by Hunter and Schmidt (1990), conclude that not one predictor conclusively generalized in terms of predictive validity across samples. However, a number of variables had

generalisable validity moderated by various factors, including decade of the particular study, aircraft type, arm of service and nationality. The variables that had generalisable validity (with mean sample-weighted correlations indicated) included job sample (0,34), gross dexterity (0,32), mechanical ability (0,29), reaction time (0,28), biodata inventory (0,27), aviation and general information (0,22), perceptual speed (0,20), spatial ability (0,19) and quantitative ability (0,11). Validities that could not be generalized across samples were verbal ability (0,12), fine dexterity (0,10), age (-0,10), education (0,06) and personality (0,10).

Similar results are reported by Martinussen (1996) in a meta-analysis of 66 independent samples from 50 studies (combined N=17900) from 11 nations, also using the Hunter and Schmidt (1990) meta-analysis method. She found the best predictors of pilot performance to be - with mean corrected validities indicated - a combination of cognitive and psychomotor tests (0,37), previous training experience (0,30), cognitive abilities (0,24), psychomotor/information-processing abilities (0,24), aviation information (0,24) and biographical inventories (0,23). Similar to the findings of Hunter and Burke (1994), certain factors were found to have low mean validities, including personality (0,14), intelligence or *g* (0,16) and academic tests (0,15).

In a smaller follow-up meta-analysis of 4 studies (combined N=973), again using the Hunter and Schmidt (1990) method, Martinussen and Torjussen (1998) found that the best predictors of success in pilot training were instrument comprehension (0,29), mechanical principles (0,23) and aviation information (0,22).

A general conclusion can be made after review of the literature on predictors of pilot training success. Predictors vary across time frames, technology and development in the nature of the task of the military pilot. This underscores the importance of validation within the particular context of use of a selection battery. As Huysamen (1994) caveats, "it is therefore more appropriate to refer to the validity of a test for a particular application than to speak of the validity of a test" (p. 31).

Currently, there is general consensus that the ability to predict pilot success leaves much to be desired. Obtained multiple correlations are still low (Damos, 1996). Much of this relates to the choice of criterion, and unique problems associated with

pilot selection research, such as small selection ratios and severe restriction of range (Burke et al., 1997; Carretta, 1992a; Hilton & Dolgin, 1991). Fortunately, recent optimism about predictiveness has surfaced, where it is proposed that more valid and reliable criterion measures be developed, and that research into new models of personality is conducted (Damos, 1996).

In line with these findings presented, the SAAF continuously attempts to refine its pilot selection batteries (Aspeling, 1980; Croucamp & Bolton, 2002; Smit & Bielfeld, 2001). Therefore, if so much rests upon the quality of the selection decisions made in selecting SAAF pilots, it is critical that the selection battery in question be validated.

Research question

To what extent is the existing psychometric assessment battery for selecting a group of South African Air Force pilots a valid and fair predictor of training performance?

The aim of the present study is firstly to compile a battery of tests that can predict pilot training performance as measured by results for officer's formative training, ground school training and practical flight training. The following constructs are considered for this purpose: fluid intelligence, spatial ability, general intellectual ability, conceptualization ability, memory, mathematical ability, observational ability and psychomotor coordination. Secondly, this study investigates whether the current selection battery displays predictive bias with respect to gender and population groups.

Hypotheses

The aims of this study can therefore be stated in the form of testable hypotheses:

1. The pilot training performance (FLY, GROUND & FORM) of SAAF pilots can be predicted by means of measures of general/fluid intelligence (RAVENS), spatial ability (BLOX), general intellectual ability (AAT), conceptualization ability (SPX100), memory (SPX200), mathematical ability (SPX302), observational ability (SPX400) and psychomotor coordination (SPX2600).

2. The results of the selection battery are not predictively biased against specific population groups when used for selecting new trainee pilots for the SAAF.

METHOD

Sample

The sample consists of recently qualified SAAF pilots (N=107), i.e. they completed officer's formative training, ground school training and practical flight training successfully, from 1997 to 2002. Their ranks upon entering training ranged from candidate-officer to major, where most (85,1%) resorted in the former category. In terms of gender, 101 of the pilots were males and six were females. The pilots were all under the age of 25 upon entering the training programme. All of the pilots had completed at least matric. The distribution of the gender and ethnic groups in the sample is shown in Table 1.

TABLE 1: DISTRIBUTION OF SAMPLE STATISTICS

Population Groups	Male	Female		
	N	N	Total N	%
African	7	1	8	7,5%
Coloured	6	0	6	5,6%
Indian/Asian	5	0	5	4,7%
White	83	5	88	82,2%
Total	101	6	107	
Percentage	94,4%	5,6%	100	

Data analysis

The statistical techniques included descriptive statistics, Pearson Product-Moment Correlation Analysis, and Stepwise Multiple Regression Analysis (Tabachnick & Fidell, 1989). The analysis of the data was planned to be concluded with an investigation of the predictive bias of the selection battery by means of regression based procedures (Arvey & Sackett, 1993). The Statistical Package for the Social

Sciences (SPSS) was used for all analyses, where an alpha level of 0,05 was used for the determination of significance levels of all tests, unless stated otherwise (SPSS, 1999). Using the tables of Cohen (1988), statistical power for this study was estimated at 0,87 (N=107; estimated effect size = 0,30).

Measuring instruments

Criterion measures:

The criterion for this study was subjects' performance during the total pilot training process. Therefore, instructors' ratings of practical flight performance (FLIGHT), training grades for ground school flight training (GROUND) and scores on officers' formative training (FORM) were considered as measures of the dependent variable. In this regard the recommendation of Damos (1996) - that more criterion measures be included in pilot selection validation studies - was followed. Unfortunately, evidence on the reliability and validity of the criteria were not available at the time of the study. This is a common weakness of pilot validation studies (Hunter & Burke, 1994; Martinussen, 1996). In their meta-analysis of published studies, Hunter and Burke (1994) found the most common criterion to be a dichotomization of training outcome into pass-fail categories. It could be argued that this might be the reason why pilot training performance criteria are so seldom discussed in pilot selection validation studies. One Norwegian study estimated the reliability of its criterion (theoretical tests and pass/fail measure) to be 0,90 (Martinussen & Torjussen, 1998).

Predictor measures:

Raven's Advanced Progressive Matrices (RAVENS)

The Raven's Advanced Progressive Matrices measures the concept of fluid intelligence advanced by Cattell (Raven & Court, 1998). As a measure of general intellectual ability or *g*, the RAVENS is often used in the selection of staff for high-level technical positions. Reliability, as well as construct and predictive validity of the instrument, has been established in numerous studies (Bors & Stokes, 1998). Martinussen and Torjussen (1998) report corrected mean validities of the Raven's in their meta-analysis of pilot selection studies of 0,16.

Blox (BLOX)

The BLOX Test is a test of spatial ability or, more specifically, spatial relations, orientation and visualisation. Reliability estimates for the BLOX could not be obtained, but various studies illustrate adequate construct and predictive validity, mostly in the engineering and trade environment (Lombard, 1980).

Academic Aptitude Test (AAT)

The Academic Aptitude Tests (Minnie & Paul, 1993) is a battery of nine tests that measures various aspects of intelligence. One of the subtests used in this study was the AAT1, which provides an indication of the non-verbal reasoning ability of a person at Grade 12 level. This is essentially a measure of general intellectual ability (*g* or intelligence). The AAT2 was also used in this study and it provides a measure of a person's verbal reasoning ability.

The reliability (Kuder-Richardson 20) of the AAT1 has been reported as 0,88 and for the AAT2, 0,79. Evidence of predictive validity is limited to the ability to predict academic success in different school subjects (Minnie & Paul, 1993).

Situation Specific Evaluation Expert (SPEEX)

A number of subtests of the Situation Specific Evaluation Expert (SPEEX) (Erasmus, 2002) system were included in the battery. The choice of subtests used in the battery resulted from a job profiling index (JPI) that was completed with the aim of identifying the necessary competencies of a military pilot. The subtests included conceptualisation, memory, advanced calculations, observance, and hand-eye/psychomotor coordination. The constructs measured by these tests are self-explanatory, except for conceptualisation and observance. Conceptualisation is similar to deductive reasoning ability, whereas observance refers to the potential or ability for detail observation. According to the test publisher, the SPEEX is "an assessment tool that guarantees internal reliabilities of 0,75 and higher per dimension" (Erasmus, 2002). No evidence of predictive validity for the SPEEX subtests in the military setting could be obtained. However, the SPEEX tests were based on the Potential Index Battery, which has shown evidence of both construct

and predictive validity in the educational and corporate environment (Kriel, 1999; Schaap, 2000).

It is important to note that, in this study, the predictive validity of the AAT and SPEEX scores were evaluated separately from that of the Ravens and Blox scores. This was due to the fact that the former tests were included in the selection process only during the last two yearly selection cycles and, therefore, limited data is available on these instruments. The SAAF has also been evaluating the Vienna Test System (Schuhfried, 2003) during this time, but the results will be excluded from the analyses since the interpretation of these scores is still a subject of debate.

Procedure

The psychometric test scores of all qualified pilots, collected during their selection for the pilot training programme, from 1997 to 2002 were retrieved from the assessment database. The psychometric test scores were combined with the training evaluation scores that were achieved after completion of their training period, and subsequently screened for inadequate data. Cases that had missing data on the primary criterion of flight training evaluation score were excluded from the study. From the above, it can be gathered that this is a predictive criterion-related validation study, although the author shares the view of Schmitt and Chan (1998), whom are of the opinion that the traditional distinction between concurrent and predictive validation studies tends to be simplistic.

RESULTS

Statistical (stepwise) regression was employed to develop a subset of predictors that is useful in predicting pilot training performance, and to eliminate those predictors that do not provide additional prediction to the predictors already in the equation (Tabachnick & Fidell, 1989). Analyses were performed using SPSS REGRESSION and SPSS DESCRIPTIVES was used for the evaluation of assumptions underlying the statistical techniques employed.

The results of the evaluation of assumptions led to transformations of the variables to reduce skewness and improve normality, linearity, and homoscedasticity of residuals. Inverse square root transformations were used on BLOX, RAVENS and GROUND, and SPX302 scores, where inverse log transformation was performed on SPX100 scores, and a reflect-and-inverse transformation was performed on SPX2600 scores. In most cases skewness was reduced with transformation, but normality was not significantly improved as judged by the respective Kolmogorov-Smirnov test statistics, which tests the hypothesis that a sample comes from a normal distribution. Therefore, transformations were not retained due to its consequent complication of interpretability of results. Besides, multiple regression analysis is believed to be fairly robust against moderate violations of the assumption of normality resulting from skewness (Tabachnick & Fidell, 1989). With the use of a $p<0,05$ criterion for Mahalanobis distance no outliers among the cases were identified. A few cases had missing data, which were deleted pairwise.

Correlations between predictor measures and criteria

The predictor constructs are seen as relatively stable sets of individual behaviour, which should lead to superiour performance on the different dimensions of pilot training success. Based on the survey of literature and reasoning followed, it was expected that the first hypothesis would be supported; in other words it was expected that intercorrelations between predictors and criteria would be statistically significant ($p<0,05$).

Table 2 depicts correlations (Pearson) between the nine predictor measure scores and the three measures of pilot training success.

TABLE 2: CORRELATIONS (PEARSON) BETWEEN PREDICTOR VARIABLES AND PILOT TRAINING PERFORMANCE

		1	2	3	4	5	6	7	8	9
Flight training performance (FLIGHT)	*r*	0,248*	0,336**	0,592**	0,197	0,304	-0,150	0,118	0,111	-0,246
	Sig. (2-tailed)	0,015	0,001	0,001	0,344	0,091	0,421	0,520	0,545	0,198
	N	96	97	26	25	32	31	32	32	29
Ground school training performance (GROUND)	*r*	0,195*(marg)	0,138	0,397*	0,337	0,211	0,207	0,310	0,208	-0,024
	Sig. (2-tailed)	0,056	0,177	0,045	0,100	0,247	0,265	0,084	0,254	0,901
	N	96	97	26	25	32	31	32	32	29
Officers' formative training results (FORM)	*r*	0,216*	0,033	0,161	0,127	0,103	-0,248	0,143	0,003	-0,186
	Sig. (2-tailed)	0,040	0,758	0,453	0,565	0,587	0,195	0,452	0,989	0,352
	N	90	91	24	23	30	29	30	30	27
1. Fluid intelligence (RAVENS)	*r*	1								
	Sig. (2-tailed)	.								
	N	96								
2. Spatial ability (BLOX)		0,415**	1							
	Sig. (2-tailed)	0,000	.							
	N	96	97							
3. General intellectual ability (*g*) (AAT1)	*r*	0,314	0,454*	1						
	Sig. (2-tailed)	0,135	0,026	.						
	N	24	24	26						
4. Verbal reasoning ability (AAT2)	*r*	0,274	0,194	0,124	1					
	Sig. (2-tailed)	0,207	0,374	0,555	.					
	N	23	23	25	25					
5. Conceptualisation ability (SPEEX 100)	*r*	-0,024	0,002	0,109	0,222	1				
	Sig. (2-tailed)	0,898	0,993	0,596	0,287	.				
	N	30	30	26	25	32				
6. Memory (SPEEX 200)	*r*	0,074	-0,062	0,145	0,286	0,170	1			
	Sig. (2-tailed)	0,703	0,748	0,479	0,166	0,360	.			
	N	29	29	26	25	31	31			
7. Advanced calculations (SPEEX 302)	*r*	0,094	-0,112	0,000	0,608**	0,265	0,203	1		
	Sig. (2-tailed)	0,621	0,554	1,000	0,001	0,142	0,274	.		
	N	30	30	26	25	32	31	32		
8. Observance (SPEEX 400)	*r*	0,620**	0,214	0,597**	-0,013	-0,026	0,053	-0,047	1	
	Sig. (2-tailed)	0,000	0,256	0,001	0,950	0,889	0,779	0,798	.	
	N	30	30	26	25	32	31	32	32	
9. Psychomotor coordination (SPEEX 2600)	*r*	-0,171	-0,356	0,168	0,389	0,100	0,297	0,269	0,131	1
	Sig. (2-tailed)	0,394	0,068	0,443	0,066	0,606	0,125	0,159	0,500	.
	N	27	27	23	23	29	28	29	29	29

From Table 2, it can be seen that the intercorrelations between the various predictors are generally low, with a few exceptions (probability values are indicated and indicate statistical significance). Fluid intelligence (RAVENS) correlated with spatial ability (BLOX) ($r=0,415$; $p<0,001$) and with observance (SPX400) ($r=0,620$; $p<0,001$). Spatial ability (BLOX) and general intellectual ability (AAT1) were related ($r=0,454$; $p<0,05$). The latter was also associated with

observance (SPX400) (r=0,597; p<0,01). Lastly, advanced calculations (SPX302) correlated with verbal reasoning (AAT2) (r=0,608; p<0,01).

All three criteria of pilot training performance (FLIGHT, GROUND, FORM) were positively associated with fluid intelligence (RAVENS) (r=0,248; r=0,195 [marginal]; r=0,216; p<0,05). Spatial ability (BLOX) was positively associated with flight training performance (r=0,336; p<0,001), but not with ground school (GROUND) (r=0,138; p>0,05) and officers' formative training (FORM) (r=0,033; p>0,05).

Stepwise Multiple Regression results

To determine the validity of the battery to predict pilot training success, the regression of the various measures of pilot training success on the scores on the psychometric instruments was computed. Stepwise regression analyses were performed for each criterion, since they represent distinctly different aspects of the training process that were of interest to the researchers. Certain predictors were omitted from this analysis, namely general intellectual ability (AAT1), verbal reasoning (AAT2) and the SPEEX subtests (SPX100, 200, 302, 400 & 2600) due to the limited data that has accumulated over the last two selection cycles. To convincingly claim that the psychometric assessments predict pilot training performance measures, a linear composite must significantly explain variance in each of the measures of pilot training performance, all partial regression coefficients must be significant and the signs of the regression coefficients should be in the expected direction.

Table 3 indicates that only one variable was included in the regression equation for flight training performance. The predictor that delivered the largest contribution was spatial ability (BLOX). A correlation of 0,336 was obtained, that indicates that 10,4% ($0,336^2$) of the variance in the pilots' flight training performance can be explained by spatial ability as a predictor. The obtained multiple correlation is highly statistically significant, $F(1,94) = 11,976$; $p < 0,001$.

TABLE 3: STEPWISE REGRESSION: DEPENDENT VARIABLE – PILOT TRAINING PERFORMANCE (FLIGHT)

		ANALYSIS OF VARIANCE			
		Source	**df**	**Sum of Squares**	**Mean Square**
		Regression	1	257,442	257,442
Multiple R	0,336	Residual	94	2020,636	21,496
R^2	0,113	Total	95	2278,078	
Adjusted R^2	0,104				
Std.Error of Estimate	4,636	F(1,94) = 11,976; p < 0,001.			

VARIABLES IN THE EQUATION						
Independent variables		**B**	**SE B**	**βeta**	**t-value**	**P**
BLOX	(Spatial ability)	0,411	0,119	0,336	3,461	0,001
	(Constant)	63,056				

The average score on pilot flight training performance can therefore be predicted by means of the regression equation (1):

FLIGHT = 0,41BLOX + 63,05 ... (1)

The corresponding multiple correlations for the prediction of pilot success in terms of ground school training could not be computed in stepwise regression as none of the variables entered the equation according to the set criteria (probability-of-F-to-enter <= 0,05; probability-of-F-to-remove >=0,10). Hence, a standard multiple regression was run, resulting in a multiple correlation coefficient of 0,205. It was not statistically significant either, F(2,93) = 2,045; p > 0,05.

For officers' formative training the multiple correlation obtained was 0,216, which was statistically significant, F(1,88) = 4,327; p < 0,05. Fluid intelligence (RAVENS) carried the largest weight for the equation predicting success during ground school training (partial r= 0,153), as well as officer's formative training (partial r=0,216). A summary of results is given in Table 4.

TABLE 4: STEPWISE REGRESSION: DEPENDENT VARIABLE – PILOT TRAINING PERFORMANCE (ALL THREE CRITERIA)

	R	F	df	p
Flight	0,336	11,976	1;94	<0,001
Ground (Method=Enter)	0,205	2,045	2;93	>0,05
Formative	0,216	4,327	1;88	<0,05

	VARIABLES		B	βeta	R_{xy}
Flight	BLOX	:Spatial ability	0,411	0,336	0,336**
Ground (Method=Enter)	RAVENS	:Fluid intelligence	0,368	0,167	0,153
	BLOX	:Spatial ability	0,112	0,069	0,064
Formative	RAVENS	:Fluid intelligence	0,449	0,216	0,216*

*p ≤ 0,05, **p ≤ 0,01

The average score on pilot ground school training performance can not be reliably predicted since the obtained F was statistically not significant. The average score on pilot officer's formative training performance can be predicted by means of the regression equation (2):

FORM = 0,45RAVENS + 66,56 ... (2)

During the last two pilot selection cycles, both the SPEEX and AAT subtests were added to the selection procedure. Correlation statistics are reported here, since the inclusion of these variables in a regression analysis would restrict the sample size to unacceptable levels (Babbie & Mouton, 2001). The results of the correlation analysis of the SPEEX and AAT with the three criteria are depicted in Table 2. It appears that general intellectual ability (AAT1) was positively associated with both pilot flight performance (r=0,592; p<0,001) and with ground school training performance (r=0,397; p<0,05). Lastly, none of the SPEEX

subtests (conceptualization, memory, advanced calculations, observance, and hand-eye coordination) were related to any of the criteria.

As is common in most pilot selection validation studies, due to range restriction (Thorndike, 1949), obtained correlations or validities will tend to underestimate the true validities of predictors in the battery simply because the full range of ability is not present in the validation sample (Hunter & Schmidt, 1990). Unfortunately, selection data for the unselected group was not available, which is a necessary requirement for adjusting the obtained validity coefficient for restriction of range (Guilford, 1954).

The relationship between the various criterion measures is depicted in Table 5. The results were highly satisfactory in the sense of criterion convergence, since pilot flight training and ground school training were strongly correlated and highly statistically significant (r=0,424; p<0,001), thereby giving an indication that they do converge; this serves as evidence of construct validity of the criterion. On its part, officers' formative training (FORM) was not related to the other two criteria, thereby indicating that it measures aspects of training performance that are not necessarily related to the flying task.

TABLE 5: CORRELATIONS (PEARSON) BETWEEN CRITERIA OF PILOT TRAINING PERFORMANCE

		FLIGHT	GROUND	FORM
Flight	r	1	0,424**	0,051
	Sig. (2-tailed)	.	0,000	0,612
	N	108	107	100
Ground	r	0,424**	1	0,128
	Sig. (2-tailed)	0,000	.	0,204
	N	107	107	100
Formative	r	0,051	0,128	1
	Sig. (2-tailed)	0,612	0,204	.
	N	100	100	100

** Correlation is significant at the 0,01 level (2-tailed).

The final analysis relates to the fairness of the selection procedure, testing the hypothesis that the results of the selection battery are not biased against specific gender or population groups. For this purpose, the view of fairness as a lack of predictive bias was followed (APA, 2003). This view holds that predictor use can be seen as fair if a common regression line can be used to describe predictor-criterion relationships for all sub-groups of interest, i.e. group differences in regression slopes or intercepts signal predictive bias. Moderated multiple regression was planned for this purpose (Bartlett, Bobko, Mosier & Hannan, 1978), where the criterion measure is regressed on the predictor score, group membership, and an interaction term between the two. Unfortunately, severely unequal (and in some cases very small) sub-sample sizes (see Table 1) made this analysis unfeasible.

DISCUSSION

The primary aim of this study was to determine the regression of pilot training performance during flight, ground school and officers' formative training on the scores on the psychometric instruments. Individuals with higher levels of fluid intelligence, spatial ability, general intellectual ability, conceptualization ability, memory, mathematical ability, observational ability and psychomotor coordination should achieve better training scores in pilot training. The hypothesis thus stated that there is a significant relationship between pilot training performance and the predictors in the battery; this hypothesis found disparate support in this research. An analysis of the regression results (by interpreting predictor-criterion correlations as well as the various beta-coefficients) leads to the following interpretation.

From Table 2, it can be seen that the intercorrelations between the various predictors are generally low in most instances. This indicates that the battery, as a whole, measures distinctly different variables.

All three criteria of pilot training performance were significantly positively associated with fluid intelligence. Assuming the argument made earlier that fluid intelligence and general cognitive ability should be theoretically congruent to some extent, this finding supports earlier research on the prominent role of general cognitive ability (*g*) in predicting pilot training performance (Damos, 1996; Hilton &

Dolgin, 1991; Ree & Carretta, 1996). In support of this line of thinking, the measure of general intellectual ability (AAT1) was positively associated with both pilot flight and ground school training performance. The obtained association could be expected since the AAT1 also essentially measures *g*. This observed relationship mirrors the findings of Carretta and Ree (1996). As expected, spatial ability was positively associated with flight training performance, similar to the results of Carretta and Ree (1996). It confirms the assumption that spatial relations and orientation play an important part in the actual task of flying an aircraft. Interestingly, spatial ability was not related to ground school and officers' formative training; there is also no apparent theoretical link to be made between these constructs. The fact that none of the SPEEX subtests (conceptualization, memory, advanced calculations, observance, and hand-eye/psychomotor coordination) were related to any of the criteria could not be explained - most of these constructs could be expected to relate to pilot training performance - although the results should be interpreted with caution due to the small cell sample size (N=30).

Only spatial ability (BLOX) was included in the regression equation of flight training performance, probably due to the fact that it has a strong relation to the task of flying. The reason why fluid intelligence (RAVENS) was not included in the equation was probably due to collinearity with spatial ability (BLOX), as is evident from their strong positive correlation ($r=0,415$; $p<0,0001$). Consequently, the question arises as to the size of any additional variance that can be explained by the inclusion of fluid intelligence (RAVENS) in a model already containing spatial ability (BLOX). The partial correlation for the RAVENS in this model (0,126) indicated that it explained only 1,58% ($0,126^2$) of additional variance in flight training performance not yet accounted for by spatial ability. Concomitantly, analysis of collinearity diagnostics indicate that the variables in this model were multicolinear (tolerance = 0,828). Clearly, the use of both fluid intelligence and spatial ability in the equation is redundant. This finding concurs with that of Carretta and Ree (1996) when they state that specific abilities (e.g. spatial ability) are highly saturated with *g*. Hence, it also refutes that of Martinussen (1996) that the specific intelligence abilities have incremental validity over and above *g*.

Ground school training scores could not be reliably predicted. This is an unexpected finding, since, theoretically at least, the nature of the task and its corresponding trait requirements in ground school training could be expected to involve a major component of cognitive functioning, i.e. general intelligence. No explanations for this result can be suggested. Officers' formative training scores could best be predicted by fluid intelligence (RAVENS). An explanation for this finding can be taken from Thorndike (1949, 1986) and Schmidt and Hunter (1998), which stated that *g* is central in predicting training and job success across hundreds of occupations. Assuming that *g* and fluid intelligence are theoretically congruent, this explanation would also hold for the latter.

In general, the results of this study are consistent with previous research on the prediction of pilot training success in two ways, namely (a) the obtained correlations between predictors and criteria, as seen from Table 2, were relatively small and (b) the predictors that seemed to best predict pilot flight training performance, as the primary criterion, were spatial ability and fluid intelligence (Burke et al., 1997; Carretta & Ree, 1996; Damos, 1996; Hilton & Dolgin, 1991; Hunter & Burke, 1994).

It seems that the inclusion of both spatial ability and fluid intelligence in the battery was redundant, since both measures did not explain unique variance in the prediction of flight training performance. Surprisingly, the SPEEX measures of memory, mathematics, observation and psychomotor coordination were not statistically related to pilot training performance, contrary to what theory would suggest. This points to the need for the SPEEX tests in the battery to be further scrutinised for reliability and construct validity, since any flaws regarding these psychometric qualities could be expected to impede predictive validity (Huysamen, 1996).

What has been shown, however, is that the selection battery is not able to predict the training performance of SAAF pilots at a satisfactory level, since it explained only 11,98% and 3,6% of the variance in pilot flight and officers' training performance scores respectively, and no reliable prediction of ground school training scores. Seen in this light, the current selection battery leaves much to be desired. In spite of this, it is not uncommon for similar levels of prediction to be

reported in pilot selection validation studies. For instance, similar levels of prediction for a complete battery were reported in more than one study in the United States Air Force (Damos, 1996).

Interpreted differently, the results of this study show that 88% of variance in flight training performance could not be explained by the predictors. One explanation for the weak prediction of criteria relates to the criterion problem. Research on criteria in pilot selection validation studies often does not receive the same attention as do the predictors, especially with regard to adequate choice, reliability and construct validity (Burke et al., 1997). In this regard, the primary recommendation of this study is that further research be done to develop more suitable, reliable and valid criteria in pilot selection in the SAAF. This study has taken a step in the right direction by including ground school training and officers training in addition to flight training scores as measures of performance, thereby acting as additional criteria.

In terms of predictors, previous research suggests various constructs that could be measured and included in future selection batteries. For instance, one South African study has illustrated the role of reaction time, form and colour discrimination time, as well as rate of information processing in pilot success (Barkhuizen et al., 2002). One area where the predictor set also seems lacking is with psychomotor aptitudes; at face value they seem not well represented in the current battery, despite research consistently finding its inclusion useful in selection (Martinussen, 1996). Fortunately, the SAAF is currently assessing the Vienna Test system - a computerised psychomotor test system - to address this deficiency (Schuhfried, 2003). A final remark on the predictors in the battery is that initial correlative data on the AAT1 (general intellectual ability) is promising and this suggests that its inclusion should improve prediction of the current battery.

Returning to the second hypothesis regarding predictive bias of the selection battery; the computation of predictive bias is problematic in the existing sample. From Table 1, it is apparent that sub-samples are severely disproportionate with respect to race and gender, thereby making any calculation of regression equations for separate groups methodologically suspect (Tabachnick & Fidell,

1989). Mere inspection of the size of each sub-sample suggests adverse impact; in other words, although the selection procedure was uniformly applied to all groups of applicants the net result is differences in the selection of various groups (Gatewood & Feild, 1998). This raises concerns over the source of the current imbalance, i.e. is it due to selection procedures or can it be traced back to the recruitment process providing a non-representative applicant pool? In any case, the current state of affairs could constitute a *prima facie* case of discrimination and warrants scrutiny in order to strengthen the organisation's case against accusations of discrimination. Although predictive bias could not be investigated in this study, other studies (e.g. Carretta, 1997) found no evidence of predictive bias or differential prediction in pilot selection batteries with respect to minority or gender groups.

The unique contribution of this study to the SAAF lies in the finding that certain predictors seem redundant in the selection battery, and others do not appear to be predicting pilot training success very well. Certain deficiencies in the predictor construct set measured in the current battery were also pointed out. In light of the impending migration by the SAAF to the new, more modern, aircraft fleet, as well as severe budget constraints foreseen for the nearby future, the revision of the current selection battery can be expected to add significant value.

It is self-evident that there are limitations to this study. The fact that the validation study could not be planned prior to the selection process and run in conjunction with it limits the validation effort in a number of ways. For one, the absence of item-level data on the predictors and criteria limits estimates of reliability to be made, as well as subsequent judgements about psychometric suitability requirements. Secondly, absence of psychometric data from non-successful applicants makes estimates of the population statistics impossible, which is a requirement for the computation of adjustments to the validity coefficients for restriction of range and unreliability in the variables (Burke et al., 1997). Future studies should be extended to include data of non-successful candidates to facilitate the adjustments to the validity coefficient necessitated by severe restriction of range. Most pilot selection studies, for instance that of Burke et al. (1997), report substantial improvements in validity coefficients when adjusted for restriction of range and unreliability of criteria. Thirdly, the cost of the total

selection process should be tracked to enable the calculation of the return on investment made by conducting the selection process in the SAAF. The utility of a selection procedure allows for informed judgements on the cost-benefit ratio of any selection procedure (Cascio, 1993).

Another limitation relates to the unexpected findings regarding some of the predictors in the battery. It seemed that the theoretically sound linkages between the variables measured by the SPEEX subtests (conceptualization, memory, advanced calculations, observance and hand-eye/psychomotor coordination) and the pilot training performance criteria were not supported by the obtained statistical relationships. This casts doubt on the psychometric properties of these instruments in this population, or it could point to a lack of adherence to standardisation and administration requirements. Deviations in this regard could limit the reliability of the instruments and ultimately a selection battery's predictive validity (Huysamen, 1996). One solution can be taken from the American experience which has proven that transferring psychometric testing to computerised testing tends to increase reliability and validity of the selection process (Carretta, 1989, 1992b; Ree & Carretta, 1998). This could be a fruitful prospect for the current selection procedure.

In conclusion, the principles for the validation and use of personnel selection procedures (APA, 2003) warn that the results of a local validation study should be interpreted with caution, as validity coefficients may fluctuate from one sample to the next. Therefore, it is suggested that the results of this study be cross-validated in a future study. Since sufficiently sized validation samples in the SAAF take many years to accumulate, it is suggested that collaboration with similar institutions in the private and non-governmental sectors be investigated in order to share data for validation purposes (Sackett & Arvey, 1993). At the same time it must also be cautioned that a stamp-collecting approach to validation, with an exaggerated emphasis on statistical validities obtained, is undesirable (Landy, 1986). Validation is essentially a process of hypothesis testing. Therefore, it is possible that sound theorising and informed judgement, based on a thorough and methodologically sound analysis of the job and corresponding required knowledge, skills, aptitudes and other characteristics (construct validity), suggest the inclusion of measures that do not seem to statistically relate (criterion validity) to the

criterion. In this case, professional judgement should serve as sufficient evidence for its inclusion in a selection procedure (Schmitt & Chan, 1998). Suggestions for future research emanating from this study includes the analysis of criteria in pilot training selection in terms of relevancy, deficiency and contamination, as well as the incremental validity of measures of the five-factor model of personality in pilot selection.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Psychological Association. (2003). Principles for the validation and use of personnel selection procedures. (4th ed.). Washington, DC: APA.

Arvey, R.D., & Sackett, P.R. (1993). Fairness in selection: Current developments and perspectives. In N. Schmitt & W.C. Borman (Eds.), Personnel selection in organizations (pp. 171-202). San Francisco: Jossey-Bass.

Ashman, A., & Telfer, R. (1983). Personality profiles of pilots. Aviation, Space, and Environmental Medicine, 54 (10), 940-943.

Aspeling, E.G. (1980). Vlieënierskeuringstrategie vir die jare 1980 tot 1990 (Special Report U/Pers 109). Braamfontein: Nasionale Instituut vir Personeelnavorsing (WNNR).

Babbie, E., & Mouton, J. (2001). The practice of social research. Cape Town: Oxford.

Barkhuizen, W., Schepers, J.M., & Coetzee, J. (2002). Rate of information processing and reaction time of aircraft pilots and non-pilots. Journal of Industrial Psychology, 28 (2).

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 31, 233–242.

Binning, J.F., & Barrett, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. Journal of Applied Psychology, 74 (3), 478-494.

Bors, D.A., & Stokes, T.L. (1998). Raven's Advanced Progressive Matrices: Norms for first year university students. Educational and Psychological Measurement, 58(3), 382- 398.

Bourn, J. (2000). Training new pilots (Report HC 880 Session 1999-2000). London: National Audit Office, UK Ministry of Defence.

Burke, E., Hobson, C., & Linsky, C. (1997). Large sample validations of three general predictors of pilot training success. International Journal of Aviation Psychology, 7(3), 225-234.

Campbell, J.P., McCloy, R.A., Oppler, S.H., & Sager, C.E. (1993). A theory of performance. In N. Schmitt & W.C. Borman (Eds.), Personnel selection in organizations (pp. 35-70). San Francisco: Jossey-Bass.

Carretta, T.R. (1989). USAF pilot selection and classification systems. Aviation, Space, and Environmental Medicine, 60, 46-49.

Carretta, T.R. (1992a). Understanding the relations between selection factors and pilot training performance: Does the criterion make a difference? International Journal of Aviation Psychology, 2(2), 95-106.

Carretta, T.R. (1992b). Recent developments in U.S. Air Force pilot candidate selection and classification. Aviation, Space, and Environmental Medicine, 63, 1112-1114.

Carretta, T.R. (1997). Group differences on US Air Force pilot selection test. International Journal of Selection and Assessment, 5, 115-127.

Carretta, T.R. (2000). U.S. Air Force pilot selection and training methods. Aviation, Space and Environmental Medicine, 71(9), 950-956.

Carretta, T.R., Perry, D., & Ree, M.J. (1996). Prediction of situational awareness in F-15 pilots. International Journal of Aviation Psychology, 6(1), 21-41.

Carretta, T.R., & Ree, M.J. (1989). Pilot-candidate selection method: Sources of validity. International Journal of Aviation Psychology, 4(2), 103-117.

Carretta, T.R., & Ree, M.J. (1996). U.S. Air Force pilot selection tests: what is measured and what is predictive? Aviation, Space, and Environmental Medicine, 67(3), 279-283.

Cascio, W.F. (1993). Assessing the utility of selection decisions: Theoretical and practical considerations. In N. Schmitt & W.C. Borman (Eds.), Personnel selection in organizations (pp. 310-340). San Francisco: Jossey-Bass.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd ed.). Hillsdale: Erlbaum.

Cook, M. (1999). Personnel selection (3rd ed.). Chichester: Wiley.

Croucamp, Y., & Bolton, S. (2002). Pilot selection: Statistical report. Pretoria: Department of Defence (RSA), Military Psychological Institute.

Damos, D.L. (1996). Pilot selection batteries: Shortcomings and perspectives. International Journal of Aviation Psychology, 6(2), 199-209.

Erasmus, P.F. (2002). Situation Specific Evaluation Expert (SPEEX). Rant-en-Dal: Potential Index Associates.

Gatewood, R.D., & Feild, H.S. (1998). Human resource selection. (4th ed.). Orlando: Harcourt Brace.

Griffin, G.R., & Koonce, J.M. (1996). Review of psychomotor skills in pilot selection research of the U.S. military services. International Journal of Aviation Psychology, 6(2), 125-147.

Guilford, J.P. (1954). Psychometric methods. New York: Mcgraw-Hill.

Guion, R.M. (1965). Personnel testing. New York: Mcgraw-Hill.

Hilton, T.F., & Dolgin, D.L. (1991). Pilot selection in the military of the free world. In R.Gal & A.D. Mangelsdorff (Eds.), Handbook of military psychology (pp. 81-101). New York: Wiley.

Hunter, D.R., & Burke, E.F. (1994). Predicting aircraft pilot training success: A meta-analysis of published research. International Journal of Aviation Psychology, 4(4), 297-313.

Hunter, J.E., & Schmidt, F.L. (1990). Meta-analysis: Cumulating research findings across studies. Beverly Hills: Sage.

Huysamen, G.K. (1994). Methodology for the social and behavioural sciences. Halfway House: Southern.

Huysamen, G.K. (1996). Psychological measurement (3rd ed.). Pretoria: Academic.

Kriel, H. (1999). Horses for courses: Situation specific selection revisited. Pretoria: Pretoria Technikon.

Lambirth, T.T., Dolgin, D.L., Rentmeister-Bryant, H.K., & Moore, J.L. (2003). Selected personality characteristics of student naval aviators and student naval flight officers. International Journal of Aviation Psychology, 13, 415-427.

Landy, F.J. (1986). Stamp collecting versus science: Validation as hypothesis testing. American Psychologist, 41 (11), 1183-1192.

Lombard, R.B. (1980). BLOX test administrator's manual. Pretoria: CSIR.

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

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(1), 33-45.

Milkovich, G.T., & Boudreau, J.W. (1997). Human resource management (8th ed.). Chicago: Irwin.

Minnie, R., & Paul, V.H. (1993). Manual for the Academic Aptitude Test (Standard 10) (AAT). Pretoria: Human Sciences Research Council.

Plug, C., Meyer, W.F., Louw, D.A., & Gouws, L.A. (1989). Psigologie woordeboek. Johannesburg: Lexicon.

Raven, J.C., & Court, J.H. (1998). Manual for the Ravens Advanced Progressive Matrices. London: H.K. Kewis.

Ree, M.J., & Carretta, T.R. (1996). Central role of *g* in military pilot selection. International Journal of Aviation Psychology, 6(2), 111-123.

Retzlaff, P.D., & Gibertini, M. (1987). Air force pilot personality: Hard data on the "right stuff". Multivariate Behavioral Research, 22, 383-399.

Republic of South Africa. (1998). Employment Equity Act (Act No. 55 of 1998). Pretoria: Government Gazette.

Rumsey, M.G., Walker, C.B., & Harris, J.H. (Eds.). (1994). Personnel selection and classification. Hillsdale: Lawrence Erlbaum.

Sackett, P.R., & Arvey, R.D. (1993). Selection in small *N* settings. In N. Schmitt & W.C. Borman (Eds.), Personnel selection in organizations (pp. 418-447). San Francisco: Jossey-Bass.

Schaap, P. (2000). The validity and reliability of the PIB (research report). Pretoria: University of Pretoria.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124(2), 262-274.

Schmitt, N., & Borman, W.C. (Eds.). (1993). Personnel selection in organizations. San Francisco: Jossey-Bass.

Schmitt, N., & Chan, D. (1998). Personnel selection: a theoretical approach. Thousand Oaks: Sage.

Schuhfried, G. (2003). The Vienna test system. Mödling, Austria: Schuhfried.

Siem, F.M. (1992). Predictive validity of an automated personality inventory for Air Force pilot selection. International Journal of Aviation Psychology, 2(4), 261-270.

Smit, C., & Bielfeld, R. (2001). Progress report of pilot selection data. Pretoria: Department of Defence, Military Psychological Institute.

Society for Industrial Psychology, Psychological Society of South Africa. (1998). Guidelines for the validation and use of assessment procedures for the workplace. Johannesburg: Author.

SPSS Inc. (1999). Base 10.0 applications guide. Chicago: SPSS.

Tabachnick, B.G., & Fidell, L.S. (1989). Using multivariate statistics (2nd ed.). New York: Harper & Row.

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

Thorndike, R.L. (1986). The role of general ability in prediction. Journal of Vocational Behavior, 29, 332-339.

Turnbull, G. (1992). A review of military pilot selection. Aviation, Space, and Environmental Medicine, 63(9), 825-830.

United States Department of Labor. (1978). Uniform guidelines on employee selection procedures. Washington, DC: Author.