EVOLUTION OF APTITUDE TESTING IN THE RAF

M. Bailey
Directorate of Recruiting and Selection (Royal Air Force)
RAF Cranwell, Sleaford
Lincolnshire NG34 8GZ, United Kingdom

Summary
This paper outlines the history of the RAF aptitude test system and the changes made to aptitude test development programmes and testing policies which have been driven by technological and psychological advances and the requirements to assess for different specialisations and be cost effective. Consideration is also given to the next generation of aptitude tests.

Introduction

The history of testing in the RAF can be traced as far back as the beginning of WWII. Before 1940, the RAF was almost entirely dependent on the unstructured interview as its main aircrew selection method. The interviews were entrusted to serving officers who had no other brief than to find the right 'types'. They were expected, without guidance on the relative merits of personality, attainment and skills, and without the technical aides to measure them, to decide who should be accepted or rejected for aircrew selection. It was said later, that if a candidate had been to the 'right school', was tall, smart and in possession of rugby boots and a bible, he was officer material. If he rode horses as well, he was pilot material!

Test Development During WWII

The limitation of the 'it takes one to know one' approach became glaringly apparent when the high incidence of pilot training failure rate - in the order of 50% - became a major issue at the start of WWII. A systematic testing procedure was consequently introduced as a permanent feature of the RAF selection system.

The first set of standard Aircrew Selection Board testing procedures included essay writing, a 15 minute Elementary Maths Test (EMT) and a 15 minute General Intelligence Test (GIT). There were three parallel forms of GIT, each with twenty verbal items. These became the first aircrew selection test battery and attempts were made to standardise the test results for ease of cross test comparison by expressing them as five letter grades based on a 1-2-4-2-1 population distribution but no cut-off scores were calculated.

These early tests were developed by Prof. F.C. Bartlett of Cambridge University at the request of the Air Ministry, which recognised the importance of objective selection testing in reducing training wastage rates. In 1941, it founded its own research unit, the Training Research Branch, to provide professional support in developing aptitude tests, other selection methods and training programmes. From this point on, all work on selection and training was centralised.

Four major developments then followed.

1. First, the RAF formally recognised the need for separate assessment of skills and personality characteristics as a result of the introduction of Flying Grading in 1942. 'Grading' entailed a short period of 12 hours flying at Elementary Flying Training in which the students' performance was recorded, assessed and analysed. D'Arcy (86) recollected that the Training Research Branch conducted a study showing grading to be a good predictor of subsequent flying standards and accident rates among student pilots. There was a clear correlation between ability at grading and speed of learning in subsequent training.

2. Attention then shifted from the general assessment of all aircrew to specific roles. In 1942, two years after its first introduction, the standard test battery was augmented by an electro-mechanical coordination test (SMA3) for pilot selection and a three part test that involved directions, tapping and morse for wireless operator selection. The addition of the coordination test increased the credibility of the aircrew test battery for pilot selection and it was the first time the battery was convincing enough to endure the periodic close scrutinies made of selection procedures. A second electro-mechanical coordination test (CVT), also devised by Cambridge University (Prof. K J Craik), was later added and brought into executive use in 1944. The wireless operator test was well accepted when its introduction led to a fall in wastage from the basic signal course.

3. During the same time period, the Training Research Branch, under the direction of Dr Parry and with the assistance of USAAF, developed a series of objective methods which were later introduced for aircrew selection in 1944. Most notable among them was a suite of 24 aptitude tests for the selection of all six aircrew categories. They were designed to reveal skill levels relevant to one or more elements of the job of flying. The combined effect of the measures resulted in higher standards of pilot cadets being sent to grading and, in due course, a startling reduction of pilot training wastage from 48% to 25%.

4. Lastly, although all tests were initially administered by orderlies, it soon became evident that specially trained staff were needed to cope with test administration, answering candidates' questions and marking a large number of test papers. A new

Paper presented at the RTO HFM Workshop on "Officer Selection", held in Monterey, USA, 9-11 November 1999, and published in RTO MP-55.

trade 'Clerk Personnel Selection', was formed and its members were subject to selection and a two-week course. This was another foundation stone in that it provided the basis for standardisation of test conditions and test score interpretation.

The RAF, starting from nothing in 1940, had developed a comprehensive and effective aptitude testing procedure by 1944. In general, tests developed in these four years were either knowledge based tests, that might be influenced by prior experience, or work sample tests resembling various aspects of flying tasks.

1944-1984 Consolidation

Between 1944 and 1984, many more experimental tests were devised and evaluated by psychologists of the Research Training Branch, which was by now known as Science 3. Some of the new tests were introduced to replace or supplement earlier ones. For instance, a three-test Fighter Controller Test Battery was introduced in 1953 and from it an Air Traffic Controller Test Battery was developed. However, by and large, no critical changes were made to the original system established during the war. The only main advance was that, by 1984, the number of aptitude tests used for selection was streamlined to 15 but they still shared a striking resemblance with those of 1944.

It was indeed a credit to the psychologists of the 1940s that they had devised tests that retained good predictive qualities for such a long period of time. However, it became difficult for their successors to devise tests with improved predictive ability because limited trialing opportunities and relatively rudimentary technology limited the scope for development. Early tests were either paper/pencil based or relied on obsolescent electro-mechanical apparatus. The test administration procedure and record keeping was labour intensive. There was therefore increasing frustration amongst the psychologists who had to collate both the manual records of test results and training data to evaluate the psychometric properties of tests and their predictive effectiveness. The process was time consuming and errors were easily made. By the late 1970s the concept of, and the need for, aptitude testing were well established, although there was a lack of financial resources.

For the first 20 years after aptitude testing was introduced, flying grading continued to be used in parallel, acting as a second selection filter. However, with the closure of the Preliminary Flying School in 1974 pilot selection relied entirely on aptitude test results. Due to changes in the flying training programmes and system, introduction of new aircraft and poorer quality candidates being attracted to the RAF in the 1970s, the pilot test battery's validities dropped considerably from the reported 0.34 in the early 60s to 0.18 with grading and 0.14 with training results. Inevitably by the late 1970s the utility of the tests, for both aircrew and controller, was widely questioned. The House of Commons Defence Committee recommended funds for further research and development and the proposal was endorsed by the Ministry Of Defence (MOD). This became the main impetus for the second generation of RAF selection tests.

Second Generation of Selection Tests

The advent of cheap micro-computer technology opened a whole new world for test development. This MOD computer-based test development project was conducted in two stages. The first provided computerised versions of existing tests currently used for selection. Migration to computerised testing was successful and computerised versions were in use at the Officers and Aircrew Selection Centre (OASC) by September 1985. The success of computerising existing selection tests was measured in terms of their distributional characteristics and reliabilities. Although differences in the means and variance were observed in some test scores, the computerised tests' reliabilities were as good as, if not better than, the original versions. The most noticeable improvement was in test-retest reliability, especially with the two co-ordination tests that were then used.

In the second stage, new tests specifically designed to be computer-based were developed and validated. Initially a task analysis based on the Fleishman Ability Requirements Procedure was carried out to identify the ability requirements of the RAF navigator role (Burke, 83). Ability domains relevant to Fleishman's taxonomy were then used as the basis for the development of new computer-based tests. A domain basically is a broad collection of similar aptitudes. The navigator study was immediately followed by a review of controller tasks and the development of computer-based tests for a new Air Traffic and Fighter Controller Test Battery (ATFCTB). Key features of it were the dynamic nature of some tests and that some tests involved multiple tasks which yielded several measures. A number of exploration analyses were carried out and scores on such multiple measure tests were analysed to develop algorithm based composite scores combining speed and accuracy or consistency in performance over several within test measures.

During the ATFCTB development, Hunter and Schmit (86) examined a total of twenty three tests and ninety associated test scores. Consideration was given to the psychometric properties of each test and associated scores, the inter-correlations between the different scores and each test score's predictive ability of pass/fail outcome at the end of the basic training course. Nineteen test scores were brought forward to the eventual regression analysis of which nine were identified as giving optimal prediction for both ATC and FC selection. The nine test scores were weighted to show their relative predictive powers, according to the beta values from the regression analysis. The summary score was simply the combination of the nine weighted

Table 1 *Current Set of Ability Domain Adopted by OASC*

Verbal Reasoning	This refers to the ability to interpret and reason with verbal information. It is the ability to identify patterns in presented information and to solve problems by combining sensible rules of thumb with a logical approach.
Numerical Reasoning	This refers to the same type of ability as Verbal Reasoning, but relates to information presented in numerical format.
Spatial Ability	This refers to the ability to form mental pictures and manipulate spatial information in the mind.
Attentional Capability	This refers to the efficiency with which an individual can deal with visual and auditory information in real time. It is related to 'capacity', a term often used by RAF operators, and 'attentional flexibility'.
Work Rate	This refers to the ability to deal with simple tasks quickly and accurately.
Psychomotor	This refers to eye-hand and eye-hand-foot coordination.

test scores. The ATFCTB predictive ability was found to be very good at 0.52.

Review of Testing Policies and R&D Programmes

Up to this stage, all of the RAF test battery structures could be perceived as being based on a "validity driven approach" to test battery development. No formal job analyses were carried out to identify actual job requirements and so define the structure of the batteries for each specific role. Test designs tended to be driven by psychological theories and the availability of adequate test delivery systems. It appears that tests thought to have the potential to predict training success were developed on an ad hoc basis and those validated were then included or excluded from the battery depending on their proven validity. Tests might be weighted depending on the beta weights resulting from regression analysis. Two potential problems can arise from this empirical approach to individual tests and test battery development.

First of all, the validity of a particular test and the weight allocated to it in a particular battery might vary as a result of different studies because validity data tends to be sample dependent. Statistics, such as validity coefficients and beta weights, are generated to explain the maximum variance in the data set and are, of course, driven by the same data set. However, validation studies are not always based on high quality data sets within which parameters remain unchanged and so, validity coefficients and regression generated test weights may not be replicated in subsequent validation studies. For example, there might be variations in trainee quality and/or training programmes. To take the ATFCTB as an example, although its validity remained good, it was found to have reduced to 0.44 in the latest validation study (Bailey, 96) and three of its test scores showed zero predictive ability while beta weights derived from regression analysis suggested that some test scores should be weighted differently.

Secondly, any battery structure might be biased by the range and quality of tests subjected to analyses. Attentional capability, for example, may have been a relevant attribute but no test of it might have been available for a validation study. Spatial ability might also have been relevant, but the test chosen to measure it might have had poor psychometric properties and consequently be rejected because only good tests contributed to the structure. In other words, aptitudes that should be measured by the battery might be inadvertently omitted, while other tests included in the battery might be measuring similar aptitudes and be reducing the "cost benefit" of the battery through duplication of effort.

In the 1990s, the RAF started to shift emphasis away from tests themselves to the aptitudes that they measure. A 'Domain Centred Framework" was adopted to conceptualise aptitude testing policies and to direct test battery development. This approach was originally introduced by Burke in the early 1990s and was later developed by Bradshaw, Hobson and Bailey. The change in testing emphasis has been discussed in detail in a paper by Bailey and Woodhead (96). In practice, any number of domains may be defined and a working set will probably evolve over time dependent on the organisational requirements. The current working set of domains adopted by the OASC is consistent with Carroll's work (93) and is outlined in table 1.

Four domain-driven task analyses followed, of which two were rational weight studies to identify aptitudes required for the pilot and navigator roles in order to define the structures of these two test batteries (Bradshaw, 93; Hobson, 95a). A rational weight study can be considered as a coarse job analysis in which

subject matter experts, individuals with a thorough knowledge of the job, are asked to evaluate the importance of each domain and the suitability of available tests. The other two were detailed functional job analyses on Air Traffic and Fighter Controllers (Bailey, 95; Bailey, 97). Each role was progressively broken down to individual job tasks at operational level. The importance of each job task was then weighted and the aptitudes required for each job task identified. As expected, the results of all four analyses showed that none of the existing batteries covered all the domains relevant to role. It was also clear that weights of individual tests did not reflect specific training requirements. The tasks analyses results, on the other hand, provided empirical data from which an ideal battery structure could be derived.

Another major step was taken in developing the "Domain Centred Framework". A new procedure has been adopted to calculate test battery composite scores. Instead of weighting individual tests, their results were converted into z scores and then the scores of tests in each domain were averaged to give a domain score. Domain scores were then weighted according to the task analysis results before being combined to give the composite score. In this way, the emphasis of the composite score is placed on domains rather than individual tests.

In order to impose a domain structure on a test battery it is first necessary to identify which tests are appropriate measures of which domain and which tests within each domain are rationally related to the specific role requirements. It was recognised that empirical analysis was particularly important because perceived wisdom regarding what a test actually measures is not always the same. Bradshaw (97a) carried out an audit on the whole RAF suite of tests and provided information about their qualities. Factor analyses were also used to identify the construct of the tests in relation to the different domains. As a result, psychometrically poor tests were removed from the suite and domain areas that were not adequately covered by the current range of tests were identified. Before these developments, there was no clear picture of what tests reliably measure within each aptitude domain. Now, we have a much better idea of what the shape and direction of our R&D effort should be.

The "Domain Centred Framework" has shown the following advantages.

1. It clearly indicates the types of ability required for training in different roles and those which should be measured at the point of selection to indicate candidate potential. The testing programmes tend to become driven by actual job demands and it is less likely that relevant aptitudes will be overlooked and/or inappropriate weightings assigned.

2. It is anticipated that a domain based composite score will be more robust and reliable because it is based on a number of domain scores, each of which is derived from tests covering a range of similar aptitudes.

3. Moreover, because abilities required to succeed in training were used as prediction criteria, batteries are less likely to be affected by changes in training syllabus and should remain in executive use for longer.

4. Lastly, because this model places the focus on ability domains, some of which might be common to more than one role, the same tests can be used for selecting different specialisations. Overall test time for RAF candidates can therefore be reduced. (In the RAF, we test candidates for several roles simultaneously and are able to offer alternative specialisations to candidates who have not scored sufficiently well or for whom training places might not be available in their first choice.)

The domain approach might not be the optimal answer to test battery development or to directing R&D effort, however, we have found it useful because it enforces structure onto test batteries. Otherwise, as Bradshaw (95) concluded '.... the structure would become a *hostage to fortune* and dependent upon the design and outcome of successive validation studies'.

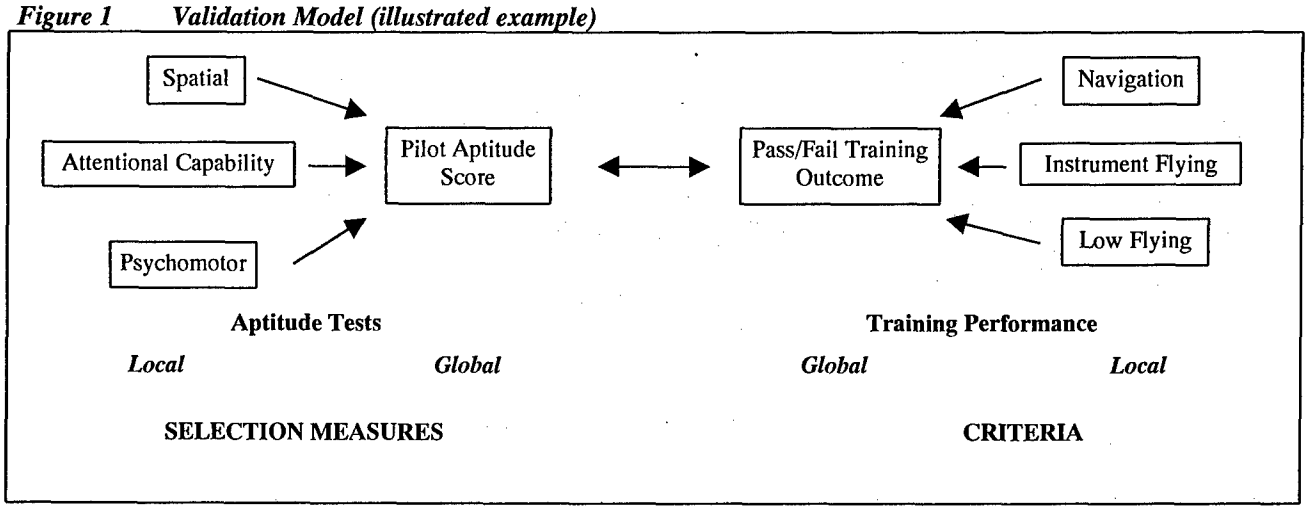
Formation of a Validation Model

In common with other organisations, RAF selection tests aim to identify individuals with the best chance of success in training. However, in order to provide an accurate indication of the effectiveness of a particular test, the choice of criterion against which it is evaluated is important. There must be a strong theoretical relationship between the selection test and the criterion used.

Traditionally, validation studies in the RAF used pass/fail training outcome as the sole criterion to evaluate the effectiveness of their tests. Pass/fail is a global criterion encompassing a number of minor local assessments of student performance. It is consistent across different training courses, making it easier to assess a test's predictive ability in general and to compare the predictive abilities of tests. However, in the 90s, training performance ratings were used more and more frequently as validation study criteria. The reasons are outlined below.

1. The non-specific nature of pass/fail is likely to attenuate the full validity of a test because it does not provide information about how well or poorly any student performed in different parts of the course, or about how good or bad was any student's overall performance. On the other hand, performance rating on a continuous scale provides quality data on overall

Figure 1 Validation Model (illustrated example)



student performance as well as progress in different elements of training. A detailed functional task analysis would identify the aptitudes required in different training areas that can be mapped onto the structure of the battery.

2. A dichotomous criterion such as pass/fail, places an upper limit on the maximum theoretical correlation coefficient that can be obtained to reflect its predictive validity. The effect is most noticeable when the dichotomy (e.g. pass rate) deviates significantly from 50%. In the 90s, this became a major concern because the pass rates of most RAF training courses were 70% or above; notably, the pilot training pass rate was in the order of 90%. If such pass rates were to remain constant or increase, the usefulness of the pass/fail criterion would decrease in proportion, despite the availability of statistical correction procedures.

The "Domain Centred Framework" approach to test battery construction and the criteria discussed so far led to the formulation of a simple validation model which is now applied to all validation studies of RAF selection methods. The concept was first introduced by Bailey (94) to validate officer qualities assessment ratings and the model was further developed by Hobson (95b, 95c) for validating the pilot aptitude test battery. It is perhaps best illustrated using the following example. If a pilot job analysis results indicated that a spatial ability test would be a useful selection tool for the instrument flying part of pilot training and a spatial test were consequently incorporated into the test battery, subsequent validity data might show that, while the new spatial test indeed related in part to pass/fail, it related best to the specific part of flying training concerned with instrument flying. Thus validation against this criterion would have provided the best estimate of that test's performance and its predictive validity would not have be overshadowed by other tests within the battery. In summary, this validation model makes a distinction between global and local selection measures and global and local criterion measures (Fig 1).

In practice, there is a danger that all predictive criteria might inter-correlate. This can be avoided to some extent by using factor analyses to categorise the different criteria in terms of ability domains with a composite score for each. The potential benefits of this validation model outweigh its potential problems. For instance, it provides a more accurate assessment of a test's validity because it refers to the specific relationship between tests and training performance. Moreover, individual tests might be predictive of certain parts of the training programme but show poor validity when considered against the global criterion. Such tests might therefore be discarded mistakenly because their true validity is hidden.

Testing in 2000s

The 1990s were characterised mostly by taking stock of our existing resources and developing a framework for steering testing policies and R&D effort. It was a period during which we consolidated and started to build upon the experience we gained in the 1980s. A range of new executive tests was introduced which included the Critical Reasoning Battery, Spatial Battery and tests of capacity; the cost effectiveness of 'test-retest' over a single test opportunity was examined (Bradshaw, 97b) and the second and third generation computer-based test systems were introduced. Each new computer system was purpose-designed, drawing on the lessons learnt from its predecessors, and took advantage of the latest available advances in information technology. This policy has in turn meant that the development potential available within each system was maximised at the time of installation. The latest system, installed in 1999, will allow us to develop tests to probe a wider range of aptitudes by exploiting system capability to deal with auditory, animated and three-dimensional presentations. The 'Domain Centred Framework' has highlighted areas requiring R&D effort such as measurement of learning

rate, ability to work with dynamic spatial information and ability to deal with a combination of aural and visual stimuli demanding more than one form of candidate response.

Since 1955, the RAF has assessed its candidates' suitability for officer training by simulating scenarios reflecting the requirements at initial officer training. The scenarios provide a platform on which candidates can reveal, first, their potential to lead a team and, second, their ability to solve practical problems using logical reasoning skills. The first is normally known as leadership qualities and the second can be considered as "effective intelligence (EI)" which is similar to Dörner's operative intelligence (79, 86). EI takes into account demands on leaders, such as having to produce innovative ideas and problem solving. It recognises the direct interaction of personality with behavioural responses, the need to measure cognitive-intellectual demands and stress resistance. This dimension is different from traditional measures of intelligence factors in that it is widely considered to be a behavioural response which can be assessed objectively during 'assessment centres' situational exercises (Dörner & Kreuzig, 83; Putz-Osterloch, 85). RAF officer candidate performance is rated by a group of trained assessors. Although the ratings are effective and predictive of training success, there are always elements of subjectivity and varying standards between assessors (inter-rater reliability) which tend to reduce assessment reliability assessment. Moreover, assessors' initial ratings of candidates' might be affected by the 'halo' effect, an overall evaluation of whether the candidate is good or bad (Bailey, 94).

A more objective, alternative way, of assessing EI and other similar kinds of performance (such as risk orientation and decision making style) might be to use computer assisted tests. If this line of research is pursued, real life scenarios, such as project planning meetings, can be simulated using computer technology to represent 'virtual' players with whom a single live candidate could interact. Tests could be developed that would be adaptive and interactive so that each candidate's responses would determine the information presented during the test and so determine the way he or she moves through each exercise. We hope to start work in this area during the new millennium.

Overview

The RAF has made great progress in its aptitude testing policies and in the quality and range of tests it uses. Most test batteries are now driven by detailed analyses of training requirements and can therefore be tailored to each role for which an aptitude testing service is provided. The information technology available to the RAF since 1985 has enabled test research and development to proceed at a rate that is very much greater than was possible before its introduction. We anticipate that further advances in information technology will enable even greater advances in aptitude testing techniques to be made in future. In addition, the psychology department has now been firmly established as the sole professional support to the OASC and is co-located with the testing system. This means easy access to data, less chance of corruption in data transmission between locations and vastly improved communication between selectors, trainers and the department. We are therefore able to be more responsive to customer requirements and provide a more efficient test development and monitoring service than was the case previously.

References:

Bailey, M. (94). An evaluation of PQs. DofR&S Psychologist Report - MOD Report(unpublished).

Bailey, M. (95). *Fighter controller task analysis study*. DofR&S Psychologist Report - MOD Report (unpublished).

Bailey, M. (96). *A cross validation study of the controller's test battery*. DofR&S Psychologist Report - MOD Report (unpublished).

Bailey, M. (97). *Air traffic controller task analysis*. DofR&S Psychologist Report - MOD Report (unpublished).

Bailey, M. & Woodhead, R. (96). Current status and future developments of RAF aircrew selection. *AGARD conference proceedings 588: Selection and training advances in aviation*, (Nov 96), p.8-1 to 8-9. NATO.

Bradshaw, J. (93). *Navigator selection process: exploring the navigator test battery*. DSc(Air) Report - MOD Report (unpublished).

Bradshaw, J. (95). *Fighter controller selection process: rational weight study*. DRA/Customer Report - MOD Report (unpublished).

Bradshaw, J. (97a). *Test audit: study & analysis 3.0/97*. Eikonica Ltd (MOD Report (unpublished) British Crown Copyright 97).

Bradshaw, J. (97b). *Aptitude retesting at OASC: study & analysis 1.0/97*. Eikonica Ltd (MOD Report (unpublished) British Crown Copyright 97).

Burke, E.F. (83). *Computer-based aptitude tests for navigators: initial results*. Ministry of Defence, London: Chief Scientist (RAF) Note for the Record - MOD Report (unpublished).

Carroll, J.B. (93). *Human cognitive abilities: a survey of factor-analytic studies*. Cambridge University Press, Cambridge.

D'Arcy, S.H.R.L. (86). Aptitude testing in the Royal Air Force 1939 – 1986. *Air Clues*, Aug, 86.

Dörner, D. (79). *Problemlösen als Informations-verarbeitung*. Stuttgart: Kohlhammer.

Dörner, D. (86). Diagnostik der Operativen Intelligenz. *Diagnostica*, **4**, 290-308.

Dörner, D. & Kreuzig, H.W. (1983). Über die Beziehung von Problemlösefähigkeiten und Maß der Intelligenz. *Psychologische Rundschau*, **4**, 185-192.

Hobson, C.J. (95a). *Pilot selection process: rational weight study*. DRA/CustomerReport - MOD Report (unpublished).

Hobson, C.J. (95b). *RAF aircrew selection: validation criteria*. DRA/CustomerReport - MOD Report (unpublished).

Hobson, C.J. (95c). RAF pilot aptitude test battery: validation against basic flying training. DRA/CustomerReport - MOD Report (unpublished).

Hunter, D. & Schmit, V.P. (86). *A computer-based test battery for selection fighter and air traffic controllers*. DSc(Air) Memo - MOD Report (unpublished).

Putz-Osterloch, W. (85). Selbstreflexion, testintelligenz und individuelle unterschiede bei der bewältigung komplexer probleme. *Sprache und Kognition*, **4**, 203-216.