# International validation of a computerised testing suite for pilot selection

AGNÈS KOKORIAN
*People technologies*
COLIN VALSLER
*Psytech Ltd.*
EUGENE BURKE
*SHL*

How do you select a pilot *and* know that the person you have selected will be a safe operator and a good investment? An impossible task – well, yes and no. Of course, any process used to select people for jobs will have a degree of uncertainty, but managing the risk involved in a hiring decision and maximising the return on the investment that follows hiring, such as training, is possible.

Up until the mid 1990s, it would be fair to say that views varied considerably on what should and should not be in a selection test battery for pilots. Indeed, the considerable variation in opinion and results presented at aviation psychology conferences is what spurred Hunter and Burke to undertake a statistical review of the results of pilot validation studies going back 50 years (Hunter and Burke, 1994 & 1995). But, that study was really a starting point rather than a conclusion in scoping the landscape to be mapped out for pilot selection. Since then, other validation studies, military and civilian, have consistently shown that pilot selection tests with a clear rational behind them do indeed provide the actuarial information that aviation organisations need to manage their pilot selection programmes more efficiently and effectively. In an era when cost and return is all the more important to the airline industry, then putting the case for valid selection tools that deliver a clear return on investment would seem to be more important than ever. That is the purpose of this paper.

**Creating a clear framework for evaluating pilot selection tools**

At the risk of stating the obvious, two views need to be aligned to realise the benefits of a valid selection system:

- the *practitioner view* that requires a clear linkage between what makes an effective pilot and what selection tools assess, and a clear statement of the financial return that will be delivered from investing in selection tools

- the *scientific view* that demands clear evidence supporting the selection system including the rational behind test development (the *construct* view) and the data supporting the predictions made from the system (the *criterion validity* view)

This paper seeks to offer a case study in addressing both views. It describes the background to the development of the tests, the psychometric evidence supporting the accuracy and consistency of the tests across various military and civilian sites and various country (language) settings, as well as data on the relationship between the tests and other tests (construct validity) and with performance criteria (criterion validity). The data reported covers a total of over 4,000 candidates across 7 sites and six countries, and includes both ab initio and PPL/CPL qualified candidates.

**The *Pil*ot *Apt*itude (PILAPT) system**

The development of the PILAPT computer-based system grew out of the meta-analysis reported by Hunter and Burke, and the design principles have been described by Burke, Kitching and Valsler (1994). In summary, these design principles were as follows:

- that the test designs be based on clearly understood measures of individual differences that research has shown are relevant to pilot performance, either in training or in operations. As such, PILAPT had to cover both handling skills (as required in ab initio training) and CRM competencies (such as situational awareness and capacity).

- that the test designs should assume no prior knowledge of flying, but should have links to key pilot performance factors that are intuitive to both candidates and users.

- that the test designs should allow for practice to avoid the influence of prior experience of video games and give all candidates a level playing field to demonstrate their potential.

- that the overall battery should be efficient and avoid redundancy and nugatory assessments.

Design work on PILAPT began in 1994 and has continued with new tests and new scoring algorithms over the nine years since. Beginning with ab initio selection for the Royal Air Force (RAF) University Air Squadrons, PILAPT has been evaluated through data provided by air forces in Chile, Denmark, Portugal and Sweden, as well as civilian airlines and training schools in the UK, Europe and Asia.

PILAPT is a fully automated test delivery system built on the TEKS technology developed by Psytech Ltd. The system caters for all aspects of the testing process from candidate log on including the capture of biographical data, instructions, test administration, test scoring, analysis of candidate performance, reporting, and data transfer to other systems. The system has crash recovery and networking capabilities.

The PILAPT battery of tests developed to date includes:

- Hands (10 minutes) – the ability to process oral (verbal) rules to execute a visual task quickly and accurately – related to absorbing and using oral (e.g. radio information) under pressure

- Sequences (8 minutes) – the ability to scan information quickly and accurately in order to find critical information – related to maintaining effective instrument scan

- Patterns (10 minutes) – the ability to ignore distracting information in order to make quick and accurate decisions under time pressure – related to maintaining focus on critical information when confronted with ambiguous situations and pressure

- Concentration (8 minutes) - the ability to maintain focus on a primary task when the conditions for that task are constantly changing – related to maintaining situational awareness

- Views (20 minutes) – a two part test of the ability to visualise and understand objects and their relationships when presented in 2 and 3 dimensions and when presented from different perspectives – related to VFR and IFR

- Deviation Indicator (7 minutes) – the ability to compensate for deviations in flight parameters with a look-and-feel based on the flight path deviation indicator (FPDI) – related to basic handling skills

- Trax (5 minutes) – a pursuit tracking task requiring the candidate to work in a 3 dimensional environment – related to advanced aircraft control

In addition to the tests above, primarily driven in design by ab initio requirements and taking around an hour in total, the PILAPT battery has been extended to include a mini-test battery named Capacity targeted at experienced direct entry candidates. Capacity takes around 15 minutes to complete and comprises a primary handling task and two secondary tasks involving visual and auditory information. Tasks are administered and measured under a combination of single, dual and triple task load conditions, and the impact of increased workload on the candidate's performance is then analysed and reported using a display similar to that shown in Figure 1 below. The data shown shows average performance for Swedish fighter pilot applicants.

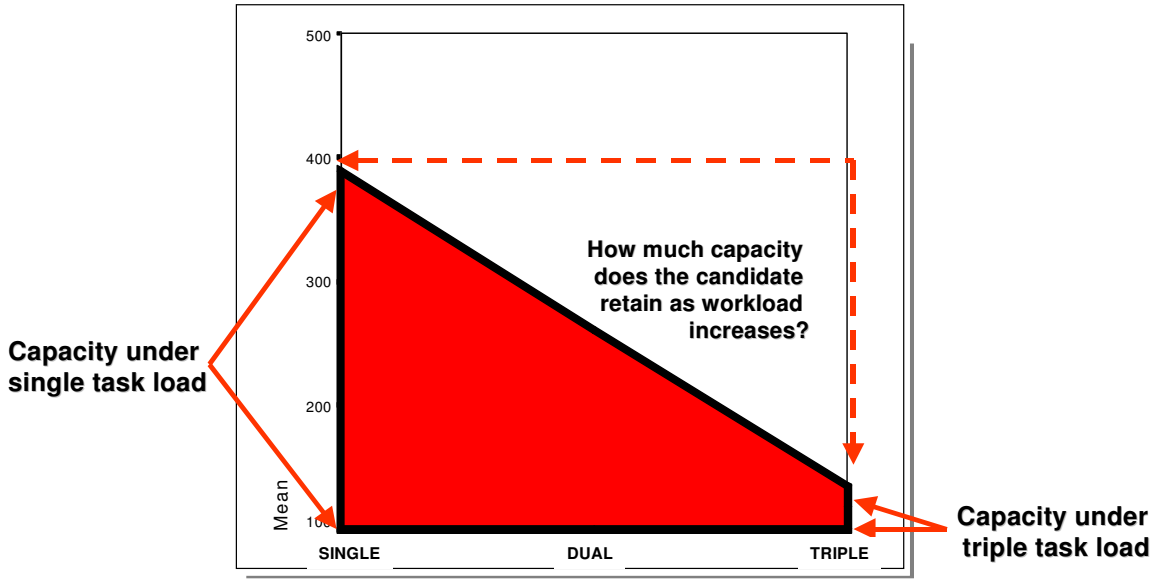


**Figure 1: Overview of what the PILAPT Capacity mini-battery measures**

**Evidence supporting PILAPT**

This section of the paper provides a summary of the data collected on the PILAPT tests to date. Given that different tests are at different stages in the development cycle, the evidence provided varies across PILAPT tests reflecting the iterative cycle of development since 1994. The evidence is presented in three parts in line with recommendations from professional bodies such as the American Psychological Association (APA), British Psychological Society (BPS) and the International Test Commission (ITC). First, evidence of test reliability (associated with accuracy and stability of scores) is presented and followed by results from studies involving other marker tests of pilot aptitude (construct validity). The last set of results summarises the evidence of the criterion validity of PILAPT in terms of predictions of training success and CRM ratings by instructors.

*Reliability*

The standard recommendation for the level of reliability required for tests used in selection is a minimum coefficient of 0.7 (this in effect states that 70% of the variation in test scores is true variation as intended in the test's design). Table 1 summarises the results of reliability (internal

consistency) analyses across various country and organisational sites. DI and Trax are not included in Table 1 as internal consistency estimates of reliability are not suitable for these tests. Data on their test-retest reliability is given below. Table 1 contains two versions of Hands, a longer 40-item version and a shorter 25-item version.

**Table 1: Reliability results for PILAPT tests across various countries and organisations**

| Source | Sample Size (N) | Hands (40 items) | Hands (25 items) | Patterns |
|---|---|---|---|---|
| 1999 Portugal (military) | 389 | 0.92 | | 0.71 |
| 2000 Portugal (military) | 162 | 0.94 | | 0.71 |
| 2000 Sweden (military) | 332 | 0.95 | | 0.71 |
| 2001 Denmark (military) | 1212 | | 0.90 | 0.69 |
| 2001 UK (civilian) | 302 | 0.96 | 0.93 | 0.79 |
| 2001 Portugal (military) | 667 | 0.93 | | 0.75 |
| 2001 Sweden (military) | 430 | 0.94 | | 0.71 |
| 2002 UK (civilian) | 638 | 0.93 | 0.88 | 0.77 |
| 2002 South America (military) | 196 | | 0.90 | 0.71 |
| 2002 Asia (civilian) | 145 | | 0.92 | 0.69 |
| Total or weighted average | 4473 | 0.94 (N=2920) | 0.90 (N=2493) | 0.72 |

In addition to these results, a test-retest (stability) study was conducted in 1995 for the RAF UAS (N=109). This study had a four month interval between test administrations and yielded reliabilities of 0.80 for DI, 0.84 for Trax and 0.77 for Hands, and an overall test-retest reliability of 0.91 for the sum of these three PILAPT test scores. More recently (2001), analysis of the internal reliability of the Concentration test for 158 Swedish military applicants yielded a reliability of 0.83. All these data clearly show PILAPT tests exceed the minimum requirement of 0.7 reliability for use in pilot selection.

*Construct validity*

This section will present two studies, both conducted in military settings, one in Denmark and one in Portugal. The Danish study was conducted with data collected in 2001 and involved four PILAPT tests – DI, Hands, Patterns and Trax – and a 15 test battery used to assess both aircrew and ATC aptitudes. Data were available across all 19 tests for a sample of 632 applicants. The content of the 15-test battery was classified according to test content in line with the classifications used by Hunter and Burke in their meta-analysis. This classification then provides a direct test of the extent to which PILAPT is measuring pilot relevant predictor constructs. The results are shown in Table 2.

**Table 2: Results for 632 Danish military applicants**

| *Test Group* | *DI* | *Hands* | *Patterns* | *Trax* | *Overall* |
|---|---|---|---|---|---|
| Mathematical Reasoning | ***.12*** | ***.31*** | ***.37*** | .06 | ***0.44*** |
| Numerical Speed & Accuracy | ***.11*** | ***.29*** | ***.25*** | .03 | ***0.35*** |
| Language | ***.18*** | ***.14*** | ***.20*** | .08 | ***0.29*** |
| General Reasoning | ***.18*** | ***.33*** | ***.51*** | .11 | ***0.57*** |
| Spatial | ***.24*** | ***.38*** | ***.38*** | ***.17*** | ***0.53*** |
| Mechanical | ***.27*** | ***.35*** | ***.40*** | ***.29*** | ***0.55*** |
| Memory | .05 | ***.23*** | ***.13*** | -.09 | ***0.27*** |

Notes:
Overall column gives the regression of the Test Group onto the 4 PILAPT tests
Correlations in bold and italicised are significant at the 0.01 level

Hunter and Burke identified the following predictor constructs as being the most consistent and substantial predictors of pilot training success: numerical reasoning, mechanical reasoning, spatial reasoning, psychomotor and simulation based tests. The Danish data set did not contain psychomotor or simulation based tests, but the results clearly show that PILAPT is tapping the other predictor constructs identified by Hunter and Burke as critical to predicting pilot training success.

The second study did contain other psychomotor and simulation tests as represented in the RAF pilot selection battery. This battery has a long pedigree (Hunter and Burke, 1987) as well as evidence supporting its validity in predicting pilot training success in the UK and Turkey across fixed and rotary wing training (Burke, Hobson and Linsky, 1997). The RAF battery comprises two tests of psychomotor ability, perceptual speed measures of aircraft recognition, as well as memory and information processing tests. However, in excess of three hours in administration time, it requires more than three times the time to administer than the full PILAPT battery. Figure 2 shows the results of regressing the overall score on the RAF battery onto a 4-test PILAPT battery requiring 35 minutes to administer. The data was obtained from 382 pilot applicants to the Royal Norwegian Air force Tested in 1997.

The figure shows that the multiple correlation between the two batteries is 0.88. This figure could be squared to provide an estimate of the variance in RAF battery scores explained by PILAPT, which would yield a percentage variance of 77% which in itself is a substantial figure. However, the straight 0.88 figure is more relevant for two reasons:

- Figure 2 shows how accurately PILAPT predicts the RAF scores and the 0.88 multiple correlation shows the degree of fit around the regression line. So, in terms of prediction, the 0.88 figure is more relevant.

- the multiple correlation can be interpreted as an alternate forms reliability measuring the equivalence in what both batteries measure – i.e. the extent to which they are both tapping similar constructs (individual differences between pilot applicants). Interpreted this way, the results suggest that there is an 88% equivalence in what is being assessed, again supporting that PILAPT is tapping predictor constructs relevant to predicting pilot success.

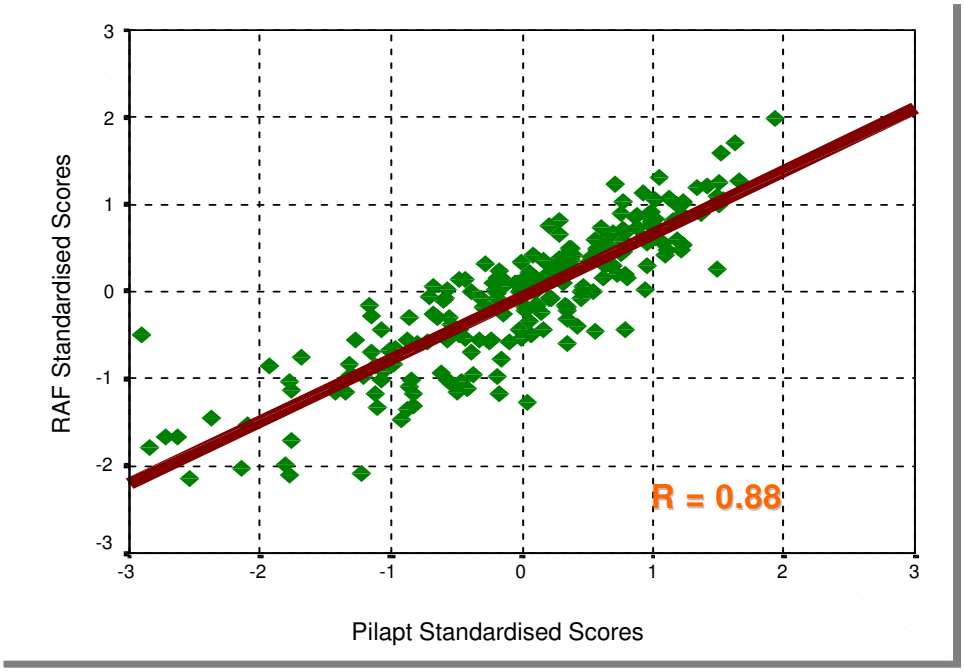


**Figure 2: Predicting RAF pilot battery scores from a short PILAPT test battery**

*Predictive (criterion) validity*

Two sets of data are described next supporting the PILAPT prediction of measures of pilot success. The first set of results are taken from a study of 165 RAF UAS students undertaking a flying training course equivalent to a PPL qualification. Table 3 presents the results in the form of correlations between three PILAPT test scores and success in training (pass versus fail). While all correlations (validity coefficients) are statistically significant, the power of the predictions shown can be gauged by using standard indices for the size of the correlations shown. That is, 0.1 represents a small validity (unlikely to yield a significant return on investment or ROI), 0.3 represents a medium validity (likely to realise a ROI) and 0.5 represents a large validity (likely to realise a significant ROI). The results clearly show that the 3-test PILAPT battery used in this instance is likely to yield a significant ROI in predicting pilot training success.

**Table 3: PILAPT predictions of PPL equivalent flying training success**

| *PILAPT Test* | *Correlation (Validity)* | *Statistical Significance* |
|---|---|---|
| DI | ***0.46*** | .001 |
| Hands | ***0.26*** | .001 |
| Trax | ***0.51*** | .001 |
| PILAPT | ***0.55*** | .001 |

The second study involved 57 commercial pilot students undergoing a UK based training course. Success data comprised a series of instructor ratings across a range of airmanship, handling, safety and CRM criteria. Across all the criteria, the average correlation between an overall 4-test PILAPT score (DI, Hands, Patterns and Trax) and instructor ratings was 0.27, statistically significant at the 0.05 level (in other words, a finding that is significantly less than chance). However, ratings are notoriously subjective and suffer from low interrater reliability with general research studies suggesting that the reliability of instructor ratings is on average around 0.5 (50% consistency). The more unreliable the criterion of success being predicted, then the lower the correlation between the predictor and success despite how well constructed and psychometrically sound the predictor is. Of course, actions can be taken to improve the quality and consistency of instructor's ratings and an improvement in the reliability of ratings to 0.7 (70% consistency) is achievable. Where such levels of consistency are obtained, then the 0.27 correlation observed in this study would be expected to increase to 0.32[1], a marginal improvement but one that yields a figure in line with general research on the validity of tests in predicting occupational outcomes.

Indeed, a correlation close to the 0.32 figure was found with a second success measure obtained following discussions with the instructors at the training school involved. They agreed that there would be some variation in their ratings subject to the students involved, but identified a particular point in the course when students transitioned to the advanced stage of the training as the crunch point in the training. They stated that this point in the course was where significant additional costs were accrued by those students requiring additional training and assessments to meet regulatory standards, and at which students were most likely to withdraw from the training programme. Performance at this stage was recorded through a number of marks awarded out of 100 for each of theory, practicals, procedures, simulation and live flying assessments. The correlation between the 4-test PILAPT battery and the overall percentage score on the advanced general handling check was 0.35, a significant and meaningful level of validity. That is, PILAPT scores administered before training would give a prediction of those most likely to succeed without additional investment in training at the crunch point in the course.

[1] This figure was obtained by using standard corrections for unreliability (attenuation) of the criterion.

**Summary**

The evidence provided has shown that a pilot test battery developed using clear design principles and targeted at differences between individuals shown to be relevant to success as a pilot does indeed predict whether people will succeed in pilot training, whether that be ab initio training or more advanced training. The evidence has also shown that the PILAPT tests have consistent psychometric properties and tap into relevant predictor constructs (aptitudes) irrespective of language and country setting, and irrespective of whether the setting is military or civilian training. In other words, PILAPT offers a benchmark of the quality of pilot applicants that is generalisable world-wide.

**How much risk is the airline accepting by not using objective assessment tools?**

Now, that is quite a claim, but one that is made for a purpose. One way in which the data presented in this paper can be used is to gauge the risk that aviation is accepting by not using substantiated tests in selecting pilots. To gauge this risk, data across all sites were used to create a world-wide benchmark of pilot aptitude. That is, data were statistically combined across all military and civilian data sets (N=4000+) to obtain an overall average (sample weighted mean) and an overall index of spread or range in pilot aptitude (a sample weighted standard deviation).

Knowing the relationship between PILAPT scores and training success, it was then possible to create a simple index of risk from 1 = high (low pilot aptitude and higher training investment) through to 5 = low (high pilot aptitude and lower training investment). In short, those falling into the high risk category of 1 were predicted to have a 36% probability of success in training, those in the mid-category of 3 a probability of 62%, and those in the low risk category of 5 a 79% probability of success (these were calculated from the various PILAPT data sets as well as results reported for other pilot test batteries such as the RAF's). Given costs of training published in the US and UK, an assumed figure of 67,500 Euros was taken as the cost per pilot trainee. Probabilities of success and estimated costs were then combined using standard finance definitions for calculating an ROI to yield some surprising figures as follows:

- Where the average level of aptitude per 100 pilot trainees falls at a benchmark of 1, then the cost per successful pilot trainee was estimated at 186,500 Euros to give a next loss of 44% for the pilot training programme. That is an ROI 0f –44%.

- Where the average level of aptitude per 100 pilot trainees falls at a benchmark of 3, then the cost per successful pilot trainee was estimated at 108,870 Euros to give a net gain of 63%.

- Where the average level of aptitude per 100 pilot trainees falls at a benchmark of 5 (obviously the preferred benchmark), then the cost per successful pilot trainee was estimated at 85,443 Euros to give a net gain of 276%.

If the reader accepts these numbers for the purposes of illustration, then what is the financial risk that organisations are presenting (in the medical sense) across the globe? Examining the benchmark scores of applicants to both military and civilian organisations can indicate this level of financial risk. If there is no screening on pilot aptitude, then one could reasonably assume that a random sample of the applicant group will be accepted into training, and the characteristics of that random sample will mirror the full applicant group. Well, Figure 3 presents the benchmark profiles of applicants to military and commercial pilot. Note that the distribution of risk is almost identical for ab initio military applicants as it is for the PPL+ applicants to the airline industry. Overall, these figures suggest that without the use of well validated pilot tests, the industry could accept up to 32% (military) and 38% (civilian) pilot hires who will require high levels of investment to achieve required skills levels, thereby creating lower ROIs from training.

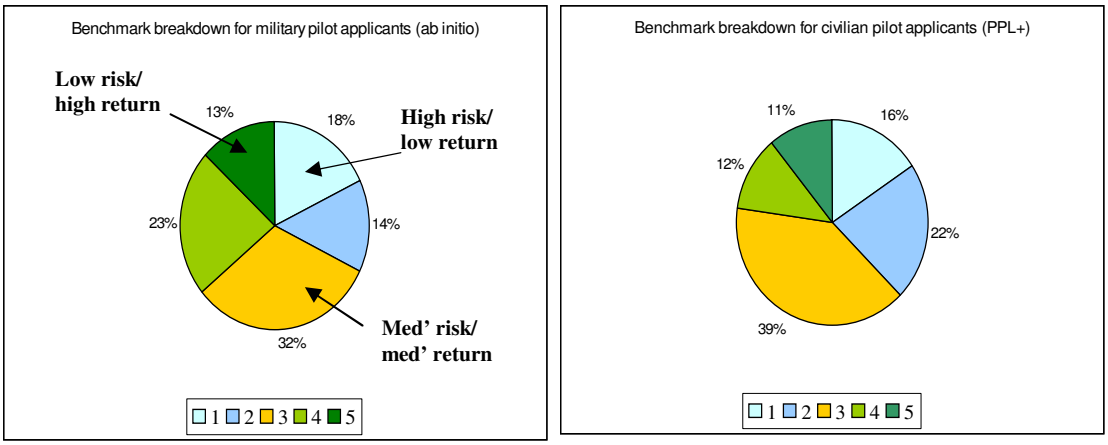


**Figure 3: Global benchmark results for applicants to military and civilian flying**

Let us make a simple logical extension. The figures shown above also indicate the quality of pilot entering the operational force and the long-term risks of not using objective pilot tests for screening applicants. That is, even if a candidate does pass training but has a low pilot aptitude, then the chances of them presenting as a risk at later stages in their career may be high. Examples include transitioning to new aircraft types and handling emergencies. Taking a systemic view of the human resource management in the aviation industry, then the importance of valid pilot tests at the point of initial hiring should, we hope, have been illustrated. And for those who would argue that flight logs provide an adequate indicator of experience, let us share a view shared with us by an experienced pilot and civilian trainer who commented to us that, every time he has shown his log book to any aircraft he has flown, he has yet to find that aircraft to provide him with an endorsement of his fitness to fly it. Quantity as shown by a flight log does not provide evidence of quality, such as capacity to handle the cockpit environment, a point clearly demonstrated over a decade ago by the work of Stead (1991) in his analysis of factors underpinning the success of Qantas pilot trainees.

**In conclusion …**

We started this paper with a question: *How do you select a pilot and know that the person you have selected will be a safe operator and a good investment?* We hope we have gone some way to answering that question and in providing a clear description of the evidence a pilot selection battery should provide to have confidence in the ROI expected back from the investment in pilot selection and the people selected with them.

**References**

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

Burke, E., Kitching, A., and Valsler, C. (1994). Computer-based assessment and the construction of valid aviator selection tests. In N. Johnston, R. Fuller, and N. McDonald (Eds.). *Aviation Psychology: Training and Selection.* Cambridge: Avebury.

Hunter, D. R., and Burke, E. (1987). Computer-based testing in the Royal Air Force. *Behavior Research Methods, Instruments, & Computers*, **19**, 243-245.

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

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

Stead, G. (1991). A validation study of the Qantas pilot selection process. In E. Farmer (Ed.) *Human Resource Management in Aviation*. Aldershot: Avebury Technical.