An Australian Trial of the British Army Recruit Battery

J.E. Greig and S.H. Bongers
Directorate of Psychology - Air Force
Suite 4, 42-52 Mawson Place, Mawson, ACT, Australia 2607

This paper reports an Australian trial of the British Army Recruit Battery (BARB), an innovative set of computer administered tests designed to provide measures of cognitive abilities. The trial was made possible because Australia, along with Canada, New Zealand, the United Kingdom and the United States of America participate in a non-atomic military research agreement called The Technical Cooperation Program (TTCP).

BARB was developed by the Human Assessment Laboratory of the University of Plymouth in collaboration with the former Army Personnel Research Establishment[2] (Kitson & Blshaw, 1996). The battery was introduced to British Army Careers Information Offices (ACIOs) on 17 February 1992, and was used on a trial basis within the South East and South West recruiting regions until its nationwide implementation on 20 July 1992.

As introduced to the ACIOs, BARB comprised seven tests, with provision to repeat one of the tests comprising the battery, or to add additional tests for evaluation should there be a requirement to do this. Excepting the time-estimation task, all tests are presented in multiple-choice response format. Responses to the multiple-choice tests are adjusted for guessing.

Scores on all tests, excluding the time estimation task, are transformed to T-scores, a scale having a mean of 50 and a standard deviation of 10. The transformed scores for the six cognitive tasks are used to calculate the General Trainability Index (GTI), which is the primary selection score. The applicant's score on Test PJ, the time-estimation task, does not contribute to the GTI. As described by Tapsfield (1995):

> The GTI is a rescaled first principal component obtained from the individual test T-scores. It takes the form of a weighted average of the T-scores, with weights obtained by adjusting the component loadings to give a composite with a standard deviation of 10 An additive constant is included in the composite to bring it onto a scale with a mean of 50....

As well as recording number of items attempted, number answered correctly and the transformed T-scores, the BARB program records the latency of the applicant's responses to the cognitive test items.

In addition to providing these benefits, BARB offered promise of having significant utility value as a screening battery because the program uses elementary cognitive tasks (ECTs) written to comply with the British guidelines published by the Equal Opportunity Commission and the Commission on Racial Equality (Allender, 1996). Differences in the performance of males and females on some of the BARB tests have been reported (Tapsfield & Wright, 1993; Collis and Irvine, 1994), however these difference have been found to counterbalance so that the GTI appears to be a gender-fair measure of ability (Tapsfield & Wright, 1993). These facts noted, however, it should be added that Irvine, Dann & Anderson (1990) clearly state that no claim is made that the tests are culture-free.

---

[1] Although TTCP provides the means for collaborative research, the authors wish to acknowledge that the Sydney trial could not have taken place without the support provided by the Ministry of Defence, Headquarters Australian Defence Force and the Royal Australian Air Force.

[2] Now the Defence Evaluation and Research Agency's Centre for Human Sciences (DERA (CHS)).

An ECT (Carroll, 1993) is a task that demands only a small set of mental operations in order to achieve a particular objective. An example provided by MacLennan (1995) is a memory-span task requiring the subject to recall a series of digits in a specific order.

By means of a process described in Irvine *et al.* (1990), the BARB program *generates* its test questions and tasks. That is to say, the program constructs test items from sets of ECTs in contradistinction to drawing from pools of already-constructed items. The algorithms employ ECTs that are within the capability of most people but present these at specified levels of cognitive demand that allow their use to measure within a wide range of ability levels. Solutions depend on cognitive processes, not on high levels of educational attainment. In fact all tests comprising the battery were designed to be comprehensible to a person with an educational achievement level of an 11-12 year old (Tapsfield & Wright, 1993).

Collis and Irvine (1994) provide an overview of 17 studies to standardise and validate the BARB core tests and the Navy Personnel Series (NPS) tests.[3] *Inter alia,* the authors reported near-normal distributions, reliabilities in the range good to excellent, and validity coefficients that are useable and consistent in what they predict and what they do not predict (Collis & Irvine, 1994, page 72). Studies with the computer administered BARB have reported a generally high level of correlation between initial tests scores and retest scores (Tapsfield, 1993). Principal component factor analyses have consistently reported single factor solutions with moderate to high component loadings (Tapsfield, 1993, 1995; Kitson & Elshaw, 1996). We note, however, that these studies have excluded Test PJ because it does not contribute to the GTI.

Early studies have shown the predictive validity of BARB tests for performance in basic military training (Holroyd, Atherton & Wright, *1995),* and in Phase 2 of military training (Jacobs, 1996). More directly relevant to the Australian trial, however, is Kitson and Elshaw's report of a study in which, at the request of the Royal Air Force, a sample of Army recruits were administered the RAF Ground Trades Test Battery (GTTB). The authors of that report examined the relationship between scores on the GTTB and scores on BARB, and assessed the suitability of BARB as a potential replacement of the RAF pencil and paper test battery (Kitson & Elshaw, 1996).

Kitson and Elshaw (1996) report that the Army sample of 422 recruits is not dissimilar to a RAF population, although the recruits were slightly less able than RAF recruits. The authors conclude that it is safe to extrapolate the results of their study to RAF candidates. The BARB GTI was found to correlate .66 with the General Ability Index (GAI) and .52 with the Ground Technical Index (GTI). Entering the BARB scores into a multiple regression yielded an R of .70 with the GAI as the dependent variable, and an R of .58 with the Ground Technical Index.

The Sydney trial was motivated by a need to introduce computer administered testing because disestablishment of the RAAF Personnel Selection Assessor mustering had engendered serious difficulties in maintaining professional standards of test administration in RAAF Psychology Sections at Australian Defence Force Recruiting Units (ADFRUs).

The decision to introduce computer administered testing turned on a requirement to achieve both invariant test administration and accurate scoring of applicant responses to the questions and tasks used in the RAAF selection process. Computer administered test programs can meet this requirement. They can also confer the advantage of accurately and automatically storing the test and biographical data needed to answer research questions and enable the calculation of norms that allow test performances to be evaluated.

---

[3] The NPS tests are pencil and paper tests developed using ECTs and the item generative algorithms of the BARB computer program. Although the items were generated by computer, they are not delivered or scored by a computer. In a sense, however, the NPS can be thought of as the Royal Navy's version of BARB.

While these advantages are associated with all computer administered tests, the item-generative algorithms in the BARB program produce what essentially are parallel forms at each administration. That is, each test administration involves different sets of items having the same difficulty levels as the items generated for previous administrations. This is a significant advantage in a screening test because the technology allows test-

retest intervals to be shortened.

The psychometric properties of BARB have been evaluated with a Canadian Forces sample with descriptive statistics and normative tables for the two primary language groups in that Country (MacLennan, 1995). That study, however, used only three pencil and paper forms of the BARB tests, one of which (Letter Distance) is not part of the computer administered battery. The Australian study is of interest because it is the first study outside the United Kingdom that employed the computer administered battery, and because it is the first study in which the same computer administered test battery has been administered to applicants for a commission.

This report presents descriptive statistics from the trial at ADFRU-S and compares those with the British data, including the data from Kitson and Elshaw's (1996) study comparing BARB and the Royal Air Force's Ground Trades Test Battery. Intercorrelations and factor loadings are reported also.

Method

Subjects

Subjects were the 235 applicants for enlistment or commissioning in the Royal Australian Air Force who were scheduled for selection testing at the Australian Defence Force Recruiting Unit, Sydney (ADFRU-S) between 19 February and 1 May 1996. The enlistment group comprised 61 males and 45 females aged between 16 and 35 years. Most applicants for enlistment were required to meet a minimum educational standard at Year 10 level, although some groundstaff occupations are open to applicants who have completed Year 9. Those who applied for commissioning included 103 males and 26 females aged between 16 and 43 years. Applicants for commissioning were required to possess the minimum educational standard of sound achievement in English and three academic subjects at Year 12 level.

Design

The independent variables were the seven tests which comprise BARB Version AC. Dependent variables were the number of correct responses on each of the BARB tests minus a correction factor for guessing.[4] In addition to corrected scores from each tests, the dependent variables included the BARB General Trainability Index (GTI), a composite score obtained by summing the weighted corrected scores on six of the seven tests. The BARB tests are briefly described in Appendix A.

BARB was administered to all applicants after they had completed all of the tests used in the selection process for their preferred RAAF occupation. Applicants for enlistment were administered the full Groundstaff Test Battery (GTB), but only the three tests used to calculate a General Index (G Index) are described at Appendix A. Similarly, category-specific test batteries were administered to applicants for commissioned service, but only the test used to assess general ability is outlined in the appendix. category-specific test batteries were administered to applicants for commissioned service, but only the test used to assess general ability is outlined in Appendix A.

Apparatus

The BARB tests were administered at twelve test stations, each furnished with a Pentium 75 micro-computer equipped with 8 Mb of RAM and a 685 Mb hard disk drive. A standard keyboard provided test administrators with an interface for entering station identification and applicant particulars before the start of testing. Test responses were made by way of a *Microtouch* 15 inch touch screen interface. A copy of the BARB software was installed on every hard disk drive, and computers were linked to a Hewlett Packard HP5/100 server for the purpose of collecting and printing each applicant's scores. All computers were

connected by means of a twisted-pair Ethernet using RJ-45 connectors. The operating system for the BARB program was MSDOS 6.22, with Windows NT 3.51 installed on the server.

---

[4] Adjusted Score = (Number Correct - Number Wrong) / *k* - 1, where *k* is equal to the number of response alternatives.

Materials

Materials included Version AC of the BARB software, which include the ability tests and routines to score responses, transform raw scores to T-scores and calculate the GTI.

The paper and pencil tests administered were those that comprised the authorised batteries for the particular RAAF occupation. In this regard, the *specialist* battery administered to an applicant for pilot training differed from that administered to an applicant for entry to the air traffic control specialisation. However, all applicants for enlistment were administered the RAAF tests WA (word knowledge), MX (arithmetic) and C (clerical abilities); these being tests used to calculate the groundstaff general ability index (G Index). All applicants for commissioning were administered Test B42, a general ability test published by ACER but restricted for use by the Australian Defence Force.

Procedure

Two weeks before the day of testing, applicants were notified that a computer delivered test battery would be administered in addition to the standard paper and pencil tests used in the RAAF selection process. A BARB test booklet was included, and applicants were advised to read the booklet and complete the practice items before attending on the scheduled test day.

The selection batteries were administered using RAAF Psychology Service standard operating procedures. Those procedures include providing timed 'breaks' at stages of testing. After completing the relevant selection batteries, applicants were provided with a 15 minute break before the BARB administration was started. Applicants were informed that the tests about to be administered were part of a process aimed at introducing computer administered tests, and that they would not be 'screened-out' for poor performance on the battery. The applicants were advised to perform to the best of their ability because their results on the computer administered tests would be considered along with other possible compensating factors should their results on the pencil and paper tests be below the required standard.

Data from the trial was analysed using SPSS for Windows Version 6 and BMDP Version 7 software packages.

Results and Discussion

Descriptive Statistics

Complete test data sets were obtained from all 235 applicants for either enlistment or commissioning, and the first analysis was aimed at determining whether these samples were representative of their larger norming groups. Those norming groups comprise all applicants for enlistment or commissioning who have taken the same tests over the running five-year period that ended on 30 June 1995. Table 1 presents the numbers, means and standard deviations used in the initial comparisons. Table 1 also presents z values calculated after testing the significance of the differences between means. Inspection of Table 1 will show that no difference between means was statistically significant ($\alpha = .05$; two-tailed test). These results allow treating data from the applicants who were tested over the period of the trial as being representative of the norming data in terms of test performance. Viewed another way, the test performances of the applicants comprising the sample are not significantly different from those of the applicants tested over the past five-years.

Table 1

Numbers, Means and Standard Deviations for the RAAF Tests tabulated by group membership

Type of Entry	RAAF Test	Five Year Norming Group			Sydney Trial Sample			Test Statistic
		N	Mean	sd	N	Mean	sd	z
Airmen	WA	12804	17.37	4.72	106	17.38	4.53	-0.0226
	MX	12792	15.79	5.20	106	16.07	4.69	-0.6116
	C	12796	21.35	5.31	106	22.15	5.36	-1.5292
Officer	B42	8306	30.84	8.30	129	31.27	8.72	-0.5562

Table 2

Numbers, means and standard deviations for applicant groups compared on the GTI, and 95 percent confidence intervals referenced to the differences between means

Group	Applicant Group First Stated in Notes			Applicant Group Next Stated in Notes			Confidence Limits Around Difference Between Means	Effect Size
	N	Mean	s.d.	N	Mean	s.d.		
A	4394	50.76	11.17	106	55.98	7.96	$3.08 \leq 5.22 \leq 7.76$	.55 sd
B	106	55.98	7.96	129	63.47	10.28	$5.15 \leq 7.49 \leq 9.83$	.82 sd
C	45	55.93	9.53	61	56.02	8.07	$-3.48 \leq -0.09 \leq 3.30$	.01 sd
D	26	63.92	9.13	103	63.35	10.59	$-3.87 \leq 0.57 \leq 5.01$	.06 sd

Notes.

1 Group A - British Army Soldier and RAAF Airmen Entry applicants.
Group B - RAAF Airmen Entry and RAAF Commission applicants.
Group C - RAAF female and RAAF male Airmen Entry applicants.
Group D - RAAF female and RAAF male Commission applicants.
2 The confidence interval limits shown in this Table are for a 95 percent level of confidence.

Table 2 presents means, standard deviations, and confidence intervals summarising four comparisons of scores on the BARB General Training Index (GTI). The British data were provided by Tapsfield (1996). For each row of Table 2, the column labelled 'Confidence Limits' presents three values separated by the symbols representing *equal to or less than.* The value in the centre of the interval is the difference between the GTI mean scores for the comparison groups identified in Note 1 to the Table. The values on the left and right of centre are, respectively, the lower and upper limits of the interval enclosing the true difference between means for theoretical populations of experimental and control subjects. The width of the confidence intervals shown in Table 2 are calculated to allow acceptance at the 95 percent level of confidence. That is to say, we may be 95 percent confident that the true value of the difference between the GTI means of each comparison group falls between the limits set out in the relevant row.

The column labelled 'Effect Size' presents a restatement of the obtained difference between means as a

proportion of the average standard deviation of the GTI scores. On examination, the data in that column show that the GTI mean for RAAF Airmen Entry applicants was .55 of a standard deviation higher than the mean for British Army Soldier applicants, and that the mean for RAAF Commission applicants was .82 of a standard deviation higher than the mean for RAAF Airmen Entry applicants. Adopting Cohen's (1977) suggestion for describing effect sizes, these differences may be classified respectively as moderate and large effects.

The finding that the GTI mean for the Airmen Entry group is .55 of a standard deviation higher than the mean for British Army Soldier applicants is consistent with Kitson and Elshaw's (1996) observation that, although their data showed that their Army sample was not dissimilar to the Royal Air Force population, subjects in their Army sample were 'slightly less able than RAF candidates.' While a definitive explanation of these findings is not possible, it may be that some potential applicants with relatively lower levels of ability decide not to apply for entry to the Air Force because of a perception that entering that Service is more difficult.

The Shapiro and Wilk statistic (Dixon, 1992) was used to test the hypothesis that the GTI scores from the 235 applicants comprising the ADFRU-S sample are drawn from a normally distributed population having a mean and variance equal to that of the sample. The obtained statistic of 0.9855 (p. = 0.6988) indicates that the distribution of scores from the sample is not significantly assymetrical. This indication was cross-checked with a normal probability plot, and by dividing the statistics for skewness (.171) and kurtosis (-.024) by their standard errors (SE_{skew} .159; SE_{kurt} .316) and evaluating the results against the normal z distribution (Tabachnick & Fidell, 1989). All indications were consistent with that provided by the Shapiro and Wilk test statistic shown above.

The 235 GTI scores spanned nearly six standard deviations, ranging from a minimum T-score of 32 to a maximum of 91. Within the distribution, the scores from applicants for enlistment and commissioning were reasonably symmetrical about their medians, given the small sample sizes. As indicated by the confidence interval and effect size shown in Table 2, the difference between the means from the two samples proved statistically significant (t = 6.1392; d.f. 233; p. <.001, two-tailed test).

Examination of the average standard deviations for the two comparisons of sex differences presented in Table 2 shows no meaningful differences between the means for male and female applicants on the GTI. The confidence intervals comparing male and female means on the individual BARB tests showed differences between the performance of the groups on some of the tests. These differences are reported in Appendix B. These findings are consistent with those of Tapsfield and Wright (1993), where the differences on individual tests counterbalanced so that, overall, there was no sex difference on the GTI.

Intercorrelations and Factor Structure

Table 3 presents the intercorrelation matrix for data from the 235 RAAF applicants. As the positive correlations imply that all tests contributing to the matrix share something in common, Fisher's z' was calculated for each correlation coefficient. Averaging those data yielded a value of .4369 which was transformed back to obtain an average Pearson's correlation coefficient of .41 for the matrix (Guilford, 1965). This index represents an estimate of the degree to which the BARB tests share a common core.

Table 3

Intercorrelations of the BARB tests and average correlation of each test with all other tests in the battery

	T2	SA	LC	ND	RF	A2	PJ

T2	1.00						
SA	.59	1.00					
LC	.55	.49	1.00				
ND	.50	.50	.41	1.00			
RF	.37	.39	.39	.35	1.00		
A2	.55	.50	.43	.33	.35	1.00	
PJ	.21	.27	.12	.30	.33	.21	1.00
Average z'	.51	.50	.43	.50	.41	.47	.25
Average r	.47	.46	.41	.46	.39	.44	.24

Note. From data yielded by 235 applicants for enlistment or commissioning in the Royal Australian Air Force.

In addition to presenting the matrix of Pearson intercorrelations, Table 3 shows the average intercorrelation of *each test* with all other tests in the battery. Inspection of those average intercorrelations will reveal their dispersion around the average index of r = .41 for the matrix as a whole, thus providing an indication of the extent to which each test shares the common core. Test PJ clearly shows the greatest dispersion below the average intercorrelation for the matrix.

To speculate about the definition of the common core would be to go beyond our data, but Table 3 shows that, with the exception of Test PJ, the average intercorrelations of each test with all other tests in the battery do not vary greatly, reflecting that there is a reasonably high degree of homogeneity among the BARB tests that are used to calculate the GTI.

A principal components analysis of the BARB scores yielded by the Australian applicants produced a single factor using the program default that extracts only factors associated with an eigenvalue of one or higher. While this single-factor solution was consistent with other reports (Tapsfield, 1993 & 1995; Kitson & Elshaw, 1996), we considered our result to be unsatisfactory for three reasons. Firstly, it accounted for only 49.98 percent of the total variance; secondly, given a matrix of positive intercorrelations, large loadings on the first principal factor are to be expected and cannot be viewed as evidence that the unrotated solution comprises one large and general factor (Kline, 1994); thirdly, both a scree plot and the second eigenvalue of .98 justified the extraction of a second factor. Accordingly, a second analysis was run with the program default replaced by an instruction to extract and rotate two factors. Table 4 presents the data from the second analysis after Varimax rotation.

Not surprisingly, inspection of Table 4 will show that, ordered by size, the loadings on the first factor are similar to the rank-order of the average intercorrelations of each test with all other tests comprising BARB. In interpreting this factor we note that the BARB tests were constructed to cover some of Carroll's second-order psychometric constructs. Considering the five tests that strongly load the first factor, we note that T2 has been conceptualised as providing a measure of fluid intelligence or g_f (Collis & Irvine, 1994) and that SA is essentially a classification task requiring the identification of similarities and differences in word meanings. As such, SA may be thought of as a measure of crystallised intelligence or g_0. In contradistinction, ND and A2 are purported to be measures of working memory, and in this regard have been linked with the general memory factor G_m. As regards the psychometric factor associated with LC, Irvine et al. (1990) state that this

is general speed or G_s, a term used by Carroll (1993) to refer to a factor measuring speed of cognitive performance.

Table 4

Principal component loadings and communalities from analysis of the RAAF data set

BARB Test	Loading		
	Factor 1	Factor 2	h^2
T2	.8190	.1399	.6903
SA	.7845	.0141	.6156
ND	.7411	.2573	.6154
A2	.7372	.2009	.5838
LC	.6382	.4420	.6027
RF	.0137	.9104	.8290
pJ	.4394	.5912	.5426

Note. The two factor solution explains 63.99 percent of the variance.

It is not difficult to see that intelligence, defined in the psychometric tradition as g_f and g_c, is needed to succeed on tasks of the kind presented by T2 and SA. However, when considering the demands of ND and A2 and LC, the role of psychometric intelligence in successful performance is less apparent. For this reason, and at this time, we conceptualise the first factor extracted from the Australian data as a cognitive performance factor. In this regard, we are acknowledging Kline (1991) and distinguishing the g_f and g_0 factors while recognising that other first and second order factors can be important when these are required by the particular task.

The data in Table 4 show that the second factor is defined by Test RF, with Test PJ also loading strongly and Test LC loading moderately. In an attempt to interpret this factor, we note firstly that Test RF is linked with the general visualisation factor (G_V) (Kitson & Elshaw, 1996; Irvine et al., 1990). Kline (1991) points out that the visualisation factor is broader than the spatial factor and loads tests and skills in which the ability to visualise is important. Secondly, Test PJ is conceptualised as a time estimation task (Tapsfield & Wright, 1993), but the person taking the test must visualise the trajectory of an icon after is disappears from view. The ability to visualise enters into the test, and we suggest that this explains PJ's loading on the second factor. Thirdly, while the factor associated with Test LC is G_S (Irvine et al., 1990), Collis and Irvine (1994) point out that the tasks associated with LC involve speed of coding and perception. They conceptualise LC as a feature detection test "that will discriminate between those who can match features of letters mentally" and those who have "trouble detecting relevant aspects of symbols required for literacy". Against this background, we suggest that Factor II is a visualization factor.

The Relationship with RAAF Tests

Table 5 shows the multiple correlation coefficients (Rs) obtained after entering scores comprising the relevant

dependent variable and BARB scores from the Australian sample into a multiple regression program. For comparison, the RAAF data are presented along with those from Kitson and Elshaw (1996). The Kitson and Elshaw data is for the RAF General Ability Index and the RAF General Technical Index, the indexes used in the airmen entry selection process.

Note: Kyllonen and Christal (1990) report a consistent and high correlation between general reasoning ability and general working-memory capacity, and they tentatively suggest that their results may be interpreted as '...supportive of the hypothesis that working-memory capacity is primarily determined by individual differences in reasoning ability'. While Kyllonen and Christal's report is persuasive, we do not wish to speculate more than would be reasonable given the limitations of our data and analyses.

Table 5

Multiple correlation coefficients (Rs) obtained by entering particular dependent variables and BARB data into multiple regressions

Type of Index	RAF Tests		RAAF Tests			
	Airmen Entry		Airmen Entry		Officer Entry	
	Dependent Variable	BARB Mult. R	Dependent Variable	BARB Mult. R	Dependent Variable	BARB Mult. R
Gen. Abil	GAI	0.70	G Index	0.75	B42	0.72
Technical	GTI	0.58	T Index	0.43	-	-

Notes:
[1] The RAF Test Indexes GM and GTI respectively mean General Ability Index and General Technical Index.
[2] The RAF tests were administered to 428 soldier recruits (Kitson & Elshaw, 1996).

On inspection, Table 5 will show that the multiple correlation coefficients obtained using Australian data supports those obtained by Kitson and Elshaw (1996) in their study using Royal Air Force selection test indexes as the dependent variables. The table also shows that the multiple correlations of BARB scores with the different tests of general ability used in the RAAF airmen and officer selection processes are similar. As no technical index is calculated for RAAF officer entry, the relevant cell in the table is empty.

Conclusions

The data from the Sydney trial support the proposition that BARB offers promise of having significant utility value as a screening test. Although the cognitive tasks presented by the BARB tests require only functional levels of literacy, the task demands range from easy through moderate to severe. In this regard, the data from the trial show that the GTI measures across a range of abilities sufficiently wide to allow its use with applicants for either enlistment or commissioning in the Royal Australian Air Force. Importantly, the GTI proved capable of discriminating between those applicant groups. Consistent with Kitson and Elshaw's (1996) findings using the RAF's General Ability Index, the data from the Australian trial also show substantial correlations between the GTI and two measures of general ability used in the RAAF selection process. The potential value of the BARB as a screening test battery is reflected in these findings, in the near-normal distribution of the GTI, and in the finding that the GTI is gender-fair.

The BARB program allows immediate retest data to be collected by making provision to repeat one of the tests after the last test in the battery has been administered. Although any particular applicant is retested once only, the program steps through the battery so that after every seventh applicant there is an immediate retest record for every test in the battery. Although this capability was engaged during the trial, the small number of

applicants tested did not provide enough data to allow calculation of reliability coefficients. In this regard, however, test-retest coefficients for British Army applicants are reported by Tapsfield (1995). Tapsfield's data show that reliability coefficients for immediate retest administrations range from 0.74 for LC to 0.88 for RF. Lower test-retest coefficients were calculated from the data for 942 people who were readministered BARB between 28 and 35 days after initial testing, but the coefficient of 0.83 for the GTI was considered to be acceptably high.

To promote reliability, operating procedures should attempt to standardise the number of days applicants have to work on their pre-test booklets before attending an ADFRU for testing. Care is also needed to ensure that all applicants know that there are advantages to be gained by carefully working through their pre-test booklets in order to understand the nature of the tasks with which they will be tested. At the ADFRU, standardised test administration procedures should include implementing controls to ensure that applicants do not write down the letters of the alphabet for use as an aid when taking Test A2.

The need for standardised administration of our current selection tests precluded counterbalancing, and thus we have no knowledge of whether there are order effects, nor, if there are order effects, of their direction or size. This question needs to be addressed in future research. Again, BARB cannot be used to screen applicants until studies establish that its measures are valid, reliable and relevant for particular RAAF occupations. Those studies will be conducted as soon as training data can be matched with scores on the battery. In the meantime, however, studies will focus on determining test-retest reliabilities and gaining a better understanding of the tests comprising the battery. An early study will employ known measures of second-order factors as markers on which to rotate the BARB tests.

References

Allender, C. (1996). The British Army Recruit Battery (BARB): A benchmark in computer delivered item-generated selection tests. *Assessment Matters. 2,* 11 - 12.

Carroll, J.B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge: Cambridge University Press.

Collis, J.M. & Irvine, S.H. (1994). A new generation of ability tests for selection and trainin2: The Navy Personnel Series. HAL Technical Rei,ort 1-1994.University of Plymouth.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (Revised Edition). New York: Academic Press.

Dixon, W.J. (Snr Ed.) (1992). BMDP statistical software manual (3 vols.). Berkley: University of California Press.

Guilford, J.P. *(1965).* Fundamental statistics in psychology and education. New York: McGraw-Hill,

Holroyd, S.R., Atherton, R.M. & Wright, D.E. (1995). Validation of the British Army Recruit Battery against measures of performance in basic military training. Centre for Human Sciences, Report DRA/CHS/H53/CR9501 9/1.0. DRA, Famborough.

Irvine, S.H., Dann, P.L. & Anderson, J.D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology. 81,* 173-195.

Jacobs, N.R. (1996). Validation of the British Army Recruit Battery (BARB) against Phase 2 military training performance measures. Centre for Human Sciences Report PLSD/CHS/H53/CR96049/1 .0, DERA

Famborough.

Kitson, N. & Elshaw, C.C. (1996). A Comparison of the British Army Recruit Battery and the RAF Ground Trades Test Battery. Centre for Human Sciences, Report DRA/CHS/H53/CR96060/ 1.0. DRA, Famborough.

Kline, P. (1991). Intelligence: The psychometric view. London: Routledge.

Kline, P. (1994). An easy guide to factor analysis. London: Routledge.

Kyllonen, P.C. & Christal, R.E. (1990). Reasoning is (little more than) working-memory capacity?! *Intelligence. 14,* 389-433.

MacLennan, R.N. (1995). New approaches to the assessment of cogmtive ability: An evaluation of the British Army Recruit Battery (BARB). Technical Note 3/95 Ontario: CFPARU.

Nunnally, J. (1967). Psychometric theory. New York: McGraw-Hill.

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

Tapsfield, P.G.C. (1993). The British Army Recruit Battery: Test-Retest Reliability. HAL Technical Report: 5 - 1993 (APRE). University of Plymouth.

Tapsfield, P.G.C. *(1995).* The British Army Recruit Battery: 1995 Applicant Norms. HAL Technical Report: 13-1995 (DRACHS). University of Plymouth.

Tapsfield, P.G.C. (1996). Descriptive statistics and a principal components analysis of data from 4394 applicants for enlistment in the British Army. Personal communication.

Tapsfield, P.G.C. & Wright, D.E. (1993). A preliminary analysis of summary data arising from the operational use of the British Army Recruit Battery. HAL Technical Report 3-1993 (APRE). University of Plymouth.

Appendix A

The BARB Tests

Transitive Inference (T2)

The transitive inference task tests is purported to provide a measure of the psychometric factor of fluid intelligence (G_f). Applicants are required to comprehend simple sentences and to use comparatives to infer a conclusion. Applicants are presented with a statement that includes the names of two people and a comparative adjective. When the applicant has read the statement, the screen is touched and a question with two alternative answers is presented. The test runs for five minutes.

Letter Checking (Feature Detection) (LC)

The letter checking task is a test of perceptual speed and coding and is purported to measure the psychometric factor of general speed (G_s). Applicants are required to recognise alphabet letters and identify

from a set of four pairs of letters the number of pairs which are the same. A pair can be in the same case, or in upper and lower case. For each item five multiple choice responses are given on the screen, from zero pairs the same to four pairs the same. The test runs for four minutes.

Alphabet Lag (A2)

The alphabet lag task is a test which is purported to measure working memory, or general memory capacity (G_m). Applicants are required to know the order of the alphabet and must use a mental counting process to convert letters, either forward or backward, to other letters of the alphabet. Two or three letters are presented to applicants followed by a plus or minus sign with a number, between -2 and +3, indicating how many steps forward or backward in the alphabet the applicants must calculate. Having calculated the new letter set applicants touch the screen to bring up three alternative answers. The alphabet is not provided so that applicants must use a reconstructive process in memory to perform the task. Difficulties are functions of letter set sizes, plus or minus direction and number of steps. The test runs for 6 minutes.

Number Distance (ND)

The number distance task is designed to test working memory (G_m). Applicants must order three numbers in high-low sequence, hold the outcome of this calculation in memory and decide whether the highest or the lowest number is further away from the number that remains. Applicants must simply touch the box on the screen which corresponds to the correct answer number. The test runs for four minutes.

Semantic Identity (SA)

The semantic identity task is a test of word meaning and is purported to measure verbal reasoning. Applicants are required to identify from three words which two words are similar in meaning and then select on the screen the odd word, or that which is not similar in meaning. The test runs until all 60 items have been attempted or for a maximum of 6 minutes.

Rotated Symbol (Spatial Orientation) (RF)

The rotated symbol task is a test of spatial rotation and is purported to provide a measure of the psychometric factor of general visualisation (G_v). Applicants are required to compare two "F" shapes that have been rotated through right angles. The applicants must decide whether the shapes are identical or mirror images. Two pairs are presented and the applicant must therefore indicate whether one pair, both pairs or neither pair are the same. The test runs for four minutes.

Time Estimation (PJ)

This task is a test of ability to estimate time through predicting the movement of a projectile through space. Applicants are shown a moving ball on a parabolic trajectory which disappears at the apex, or half way through flight. Applicants must estimate, by touching the screen at the appropriate time, when the unseen ball will hit the ground. Applicants responses are scored in accordance with how close in time the response is, to the actual time the ball took to hit the ground. The test is comprised of 48 trials, where the ball moves with three different initial angles at each of four different speeds and each combination is repeated four times.

Test 8 (WT)

The eighth test given in the experimental window is a repeat of one of the seven BARB tests, however, the instructions provided are varied. That is, the seven tests are cycled through the experimental window with

each test having one of three instruction sequences to give a total of 18 test/instruction combinations. Instructions emphasise speed with accuracy, speed at the expense of accuracy, or accuracy at the expense of speed. The window test runs for the same time as the equivalent BARB test versions.

The GTB

Test WA

Test WA is a multiple-choice test that provides a measure of word knowledge. The test is made up of 20 questions about meaning and three questions about word relationships. Test time is five minutes.

Test MX

Test MX is a measure of basic numeracy. The test comprises 13 applied arithmetic problems, nine number-series items, and six simple calculations. Test time is 12 minutes, and the questions are presented in multiple-choice format.

Test C

Test C is designed to provide a measure of clerical aptitude. The test comprises 15 problems requiring monetary calculations, 10 problems requiring the applicant to classify, tally, and order objects, and 10 questions testing spelling ability. The spelling questions are presented in multiple-choice form. Test time is 12 minutes.

The COMITB

Test B42

Test B42 is a test of reasoning comprising a number of different item types. There are 34 questions items involving word meanings, word opposites, word similarities, and word relationships. In addition to these verbal items, there are 17 questions involving either applied arithmetic or logic, 13 questions involving pictorial similarities and relationships, and 11 items involving number series. Applicants have 55 minutes in which to attempt the 75 items.

Appendix B

Supplementary Data Tables

Table 1

Numbers, means, standard deviations and confidence intervals for the difference between means for the BARB tests tabulated by sex of applicant

	RAAF Enlistment						
	Females			Males			
Test	N	Mean	s.d.	N	Mean	s.d.	Confidence Intervals
T2	45	53.80	9.55	61	52.89	9.58	$-2.81 \leq 0.91 \leq 4.63$

SA	45	58.29	11.94	61	56.98	9.62	-2.84 ≤ 1.31 ≤ 5.46
LC	45	56.76	8.71	61	52.51	9.16	0.76 ≤ 4.25 ≤ 7.74
ND	45	51.44	6.51	61	55.67	8.94	-7.34 ≤ -4.23 ≤ -1.12
RF	45	50.87	8.55	61	54.95	8.83	-7.47 ≤ -4.08 ≤ -0.69
A2	45	52.98	8.69	61	52.85	8.00	3.10 ≤ 0.13 ≤ 3.36
PJ	45	51.79	13.29	61	57.92	7.42	10.14 ≤ -6.13 ≤ -2.12
GTI	45	55.93	9.53	61	56.02	8.07	-3.48 ≤ -0.09 ≤ 3.30

Table 2

Numbers, means, standard deviations and confidence intervals for the difference between means for the BARB tests tabulated by sex of applicant

	RAAF Commissioning						
	Females			Males			
Test	N	Mean	s.d.	N	Mean	s.d	Confidence Intervals
T2	26	60.38	9.11	103	56.90	9.80	-0.68 ≤ 3.48 ≤ 7.64
SA	26	66.58	10.28	103	64.72	11.06	-2.84 ≤ 1.86 ≤ 6.56
LC	26	60.12	9.09	103	57.91	9.66	3.10 ≤ 7.21 ≤11.32
ND	26	58.08	8.92	103	61.21	10.15	-7.87 ≤ 3.13 ≤ 1.61
RF	26	55.15	7.72	103	58.05	7.94	-6.30 ≤ -2.90 ≤ 0.50
A2	26	57.96	9.38	103	57.69	8.91	-3.60 ≤ 0.27 ≤ 4.14
PJ	26	55.15	5.40	103	58.65	5.87	-5.99 ≤ -3.50 ≤ -1.01
GTI	26	63.92	9.13	103	63.35	10.59	-3.87 ≤ 0.57 ≤ 5.01

Back to Table of Contents