



**NEW TEST FORMATS FOR THE SELECTION AND ALLOCATION OF
MILITARY PERSONNEL IN GERMANY**

Sidney H. Irvine, Inpsych Ltd. UK

Johannes Wulf, Sibylle Schambach, Thomas Kutschke, Ministry of Defence, Germany

Richard F. Walker, Consultant Systems Engineer, San Antonio, Texas, USA

PAPER DELIVERED AT
THE ANNUAL CONFERENCE OF
THE INTERNATIONAL MILITARY TESTING ASSOCIATION

**Brussels, Belgium,
26th – 29th October 2004**

NEW TEST FORMATS FOR THE SELECTION AND ALLOCATION OF MILITARY PERSONNEL IN GERMANY

Sidney H. Irvine, Inpsych Ltd. UK

Johannes Wulf, Sibylle Schambach, Thomas Kutschke, Ministry of Defence, Germany

Richard F. Walker, Consultant Systems Engineer, San Antonio, Texas, USA

Origins Scope and Measurements

The first study of item-generated military tests (*Arbeitsberichte Psychologischer Dienst der Bundeswehr Nr. 1/2000*) used paper-and-pencil analogues of tests that were also delivered by computer. The paper-and-pencil test results justified the decision to conduct a further proving trial with item-generative tests delivered, administered, scored and reported by computer. With the cooperation of officials in the Bundesministerium der Verteidigung plans were approved in August for a survey to be completed by December 16 2002 in a number of testing centres in Germany. A maximum of 900 participants was requested.

Translations of the existing English language versions of the tests were first obtained and new variants of tests were produced. These were contained in an executable file capable of running within the existing German CAT4 test platform. A protocol for administration of the German language version was devised by Johannes Wulf.

Data was collated from computer-based outfiles of all the tests, regular and experimental. In January, 2003, a progress report (Irvine, Wulf, Schambach, Kutschke and Walker, 3003) was submitted with an exhaustive analysis of outcomes. Although there was overwhelming preliminary evidence of the robustness of the tests and the system, the collection of additional data permitted a summary review of the outcomes with a sample of approximately 400 participants for whom complete data sets were available. They had completed *The Bundeswehr Experimental Test Battery* and the *Bundeswehr Entry Tests*.

The Item-Generative Tests: The Bundeswehr Experimental Test Battery

The tests used in this study are variants of the *Tests for Selection Interviews* developed by Inpsych Ltd initially for clients who wanted tests that had not been previously used for military applications. The variants and isomorphs used in this research are referred to throughout as *The Bundeswehr Experimental Test Battery* (BETB) for simplicity and to provide an acronym that is

easily identified and remembered. *This label is used for convenience only. Its use does not imply either approval or ownership by the German Ministry of Defence.*

Like others before them, these test variants in the German language had to be robust, adaptable, closed to compromise; and had to occupy no more than an hour of administration. Content had to be clearly related to transferable military functions and learning skills such as perceptual scanning, literacy and numeracy, awareness of spatial directions and working memory.

To realise these goals in the German language versions, an informed approach to test construction and the development of parallel forms was adopted. To broaden the scope of the tests, two existing prototypes were added to the original English versions. Detailed specifications about what these new test items were to be, how they were to be sequenced, and how answer keys were to be constructed were implemented in German. These were discussed fully with the programming consultant, revised subsequently, and the first forms were produced.

The BETB tests were programmed initially to generate four files: a file for the computer-delivered system; and three more files for each test in Microsoft Word: a camera-ready version of the test form in paper and pencil, a matching camera-ready form with the answer key for template scoring, and a specification file unique to each test.

Nature of the BETB Tests and their Domains

The Bundeswehr Experimental Test Battery test types are derived from item-generative principles. These are described fully in the next section. Some will be seen to be working memory dependent. The computer delivered variants place more emphasis on working memory than the paper and pencil analogues. Whatever the mode of delivery, however, each test has specific variance that is non-random. For example, *Error Detection* assumes fluent text recognition, *Number Skills* requires numeracy in the basic operations of addition, subtraction, multiplication and division. *Odds and Evens* requires concept formation and recognition of numbers in written out form (e.g. one, two etc.). *Reasoning Categories* and *Word Rules* require basic vocabulary knowledge of everyday objects with a high frequency of occurrence in speech and print. The new *Alphabet Test* assumes knowledge of the order of the letters in the German alphabet, and *The Transitive Inference Test* requires the understanding of two simple sentences involving comparatives. In short, the test literacy and required knowledge levels are at a low

threshold. The processing of the information in the most demanding of the tests can be hard even for graduates from universities,

These vocabulary and grammatical constraints do not detract from the power of the tests, nor do they present educational thresholds that are unfair to minority groups. The skills and knowledge demanded are the minimum levels to be expected of people completing a normal school curriculum. In fact, elementary school children aged 11 can complete these tests quite comfortably once they understand fully what each test is about.

Gross differences in educational opportunities (for example, conscripts whose first language is not German) that have deprived individuals of permanent literacy and numeracy skills in the language used for the test instructions will, however, be reflected in depressed performance. This is not a circumstance that the tests can guard against although every effort has been made to minimise adverse testing effects. The tasks have been chosen to be fair to minorities and, as far as nature allows, to be fair to the sexes. In short, *The Bundeswehr Experimental Test Battery* (BETB) series is held to be both gender sensitive and minority conscious.

The tests are grouped for reference. Combinations of tests in different programs ensure the relevance of the tests for any particular family of occupations. The same form of test is never produced twice. Each user can be confident that the tests are unique to that organisation.

There are tests requiring basic skills, tests based on simple sentence comprehension, tests of reasoning and tests demanding visualisation.

Throughout there is an underlying demand on attention and holding information in memory long enough to process it correctly. Working memory is the single most important predictor of military training success.

Part 2: Methods and Parallel Form Results

In general, the results are presented in summary form as far as possible; in order to make the trends and conclusions clear and precise. Much preliminary analysis was carried out in order to explore every possibility.

Operational Methods

Three parallel forms of each test were administered randomly by Bundeswehr officials to cohorts within a number of different test administration centres. Protocols were provided and software was expertly installed at these centres by Johannes Wulf. The participant interface was by mouse. The cursor was always restricted in movement, so that it could never disappear off screen; and practice in mouse use was provided. Time limits for the tests of *The Bundeswehr Experimental Test Battery* are shown in Table 2.1.

Table 2.1 Time Limits and Labels for Item-Generative Tests

Label	Title	Time
AB	Alphabet Test	(4 MINUTES)
ED	Error Detection Test	(4 MINUTES)
NF	Number Skills Fluency	(5 MINUTES)
RC	Reasoning Categories Test	(8 MINUTES)
OR	Orientation (Locations) Test	(6 MINUTES)
OE	Odds And Evens Test	(4 MINUTES)
WR	Word Rules Test	(4 MINUTES)
TI	Transitive Inference	(4 MINUTES)

The analytical methods used included the scoring of response data to adjust scores for guessing; screening of data for confounding effects; the use of multivariate analysis of variance to assess form effects; estimation of reliability; regression and correlational analyses.

Results: Form Effects

Parallel forms of tests are designed to test in participants the same qualities in the same amount and with the same dispersion. That is, each form has to be as valid, to be as reliable and to have the same difficulty as every other in use.

Descriptives Parallelism Reliability

The comprehensive Tables 2.2 and 2.3 provide information necessary to assess Form Effects for all those completing every test.

Form Means and Standard Deviations

The means and variances of the three forms of the tests are consistently close. Particularly impressive are the individual test reliabilities and the very small effect sizes. Only *Word Rules* has a marginally statistically significant effect overall, but the minimal effect size (D_{\min}) shows that it can be discounted operationally. On the other hand, *Age* and *Test Competence* are always significant, even with relatively small numbers. These effects are evaluated in the relevant sections.

Form Effect Sizes

If very large numbers are tested, any mean difference, however small, between any two forms will show a statistically significant difference. The important statistic is effect size (D) reported in unit standard deviation form, so that the effect sizes are comparable from test to test. In trials of commodities, any mean effect size of less than 0.25 is generally disregarded, although there is no hard and fast rule.

In *The Bundeswehr Experimental Test Battery*, it is doubtful if any of the observed mean form differences are of practical importance because the effect sizes of the minimum form differences are only a fraction of a standard deviation in size.

As reported in Tables 2.2 and 2.3, the effect size for the two closest of any three forms (D_{\min}) never exceeds 0.10 of a standard deviation unit. The smallest difference is 0.01. In short, the two most congruent parallel forms of each test require no further equating.

The maximum effect size (D_{\max}) for the largest observed mean difference between any pair of three forms occurs in *Word Rules* where it is 0.38 of a standard deviation unit. This is never repeated elsewhere. One must bear in mind that there were 24 paired comparisons of forms on the same sample of participants. The laws of chance dictate that at least one such result should emerge from 20 different samples. Moreover, because the standard error of estimate of a single score is always used to fix border zones, the between-form effect is less than the within-form effect of normal measurement error.

Norms calculated over a number of forms minimise these marginal differences and bring them well within the limits of the error on a single participant true score estimate. Minor form and

practice effects can largely be removed if necessary, by the use of pre-test booklets containing practice examples (Tapsfield, 1993 a, b.). Use of pre-test information satisfies the need for social equity and perceived fairness to all.

Percentage Correct and Mean Log Latency Scores

For each participant, the percentage correct and mean log latency per item measured in centiseconds are also recorded. ANOVAs were carried out on all of these results in the progress report (Irvine et al., 2003) and they show little if any effects. These results were fully reported in the progress report and need no repetition.

In short, the test forms are consonant with each other in terms of average item difficulty measured by percentage correct or average time taken to complete each item. These concordant results are further evidence of form consistency.

With hindsight, one can infer that the *Reasoning Categories* test in The Bundeswehr Experimental Test Battery is in need of investigation to discover why it has proved so difficult for conscripts in the computer-delivered version. If used, the time limit might be increased. Similarly, the USAF performance on the *Number Skills* test is much lower than might be expected by European standards. This result is perhaps more easily explained by number skills educational effects than the *Reasoning Categories* differences.

Reliabilities

Estimates of reliability (internal consistency over forms) shown in Tables 2.2 and 2.3 are lower bound because of deliberate range restriction. Only those subjects who had scores in every test were used for these calculations. *Reliabilities for tests with a short time limit range from very good to excellent in the range .81 to .96. The lower-bound estimate for the reliability of a composite derived from the individual tests of The Bundeswehr Experimental Test Battery is .97.*

Table 2.2: *The Bundeswehr Experimental Test Battery Form Effects and Reliabilities*

EDAD	N	Mean	SD	RCAD	N	Mean	SD	ORAD	N	Mean	SD	NFAD	N	Mean	SD
J				J				J				J			
FORM				FORM				FORM				FORM			
1	109	32.3	8.2	1	125	9.2	8.4	1	114	17.2	8.4	1	113	35.9	15.0
2	113	31.9	6.3	2	93	9.4	8.2	2	118	15.2	6.3	2	104	37.8	14.1
3	119	33.2	7.0	3	122	10.9	9.1	3	109	15.9	7.4	3	124	35.7	13.5
Total	341	32.5	7.2	Total	340	9.9	8.6	Total	341	16.1	7.4	Total	341	36.4	14.2
Rel.	.81			Rel.	.94			Rel.	.84			Rel.	.92		
D	0.18			D	0.17			D	0.28			D	0.08		
Max.				Max.				Max.				Max.			
D	0.06			D	0.02			D	0.10			D	0.01		
Min.				Min.				Min.				Min.			
Sig. F	.400			Sig. F	.247			Sig. F	.116			Sig. F	.478		

Table 2.3: *The Bundeswehr Experimental Test Battery Form Effects and Reliabilities*

OEAD J	N	Mean	SD	WRA DJ	N	Mean	SD	TIADJ	N	Mean	SD	ABAD J	N	Mean	SD
FORM				FORM				FORM				FORM			
1	107	27.3	11.3	1	114	30.2	10.8	1	137	12.2	6.4	1	115	92.5	32.9
2	121	29.3	11.8	2	109	26.4	12.2	2	98	12.1	7.1	2	98	90.2	27.6
3	112	27.7	13.3	3	118	29.0	12.3	3	105	11.3	7.1	3	128	93.9	29.8
Total	340	28.2	12.2	Total	341	28.6	11.9	Total	340	11.9	6.8	Total	341	92.4	30.2
Rel.	.92			Rel.	.92			Rel.	.88			Rel.	.96		
D	0.16			D	0.35			D	0.13			D	0.12		
Max.				Max.				Max.				Max.			
D	0.03			D	0.10			D	0.01			D	0.05		
Min.				Min.				Min.				Min.			
Sig. F	.429			Sig. F	.048			Sig. F	.592			Sig. F	.650		

Anchor Norms for the Computer-Delivered German Version

Given minimal form effects, it is appropriate to end this section with a table of ‘anchor’ norms. Anchor norms representative of the participant population as a whole, permit the operational use of tests while additional data is being gathered to produce definitive norm tables.

Table 2.4 *The Bundeswehr Experimental Test Battery*
DESCRIPTIVE STATISTICS AND PERCENTILE NORMS FOR A NATIONAL SAMPLE OF 386 GERMAN CONSCRIPTS

		EDADJ	RCADJ	ORADJ	NFADJ	OEADJ	WRAD	TIADJ	ABADJ
N	Valid	386.0	386.0	386.0	386.0	386.0	386.0	386.0	386.0
	Missing	.0	.0	.0	.0	.0	.0	.0	.0
Mean		33.7	12.5	17.9	39.9	32.0	31.9	13.5	100.1
Std. Error of Mean		.3	.4	.3	.6	.5	.5	.3	1.3
Median		33.3	11.6	17.4	38.5	32.0	32.2	13.0	99.8
SD		6.3	8.0	6.5	12.7	9.4	9.6	5.9	26.2
Variance		40.3	63.6	42.3	160.8	88.9	92.6	35.1	687.9
Skewness		.2	.3	.4	.5	.0	.0	.4	.2
Kurtosis		.0	-.8	.5	.0	.0	-.1	.1	-.4
Range		39.3	37.4	41.8	70.0	52.3	54.0	32.5	130.2
Minimum		13.8	.1	.2	9.5	7.3	7.3	.5	39.2
Maximum		53.0	37.6	42.0	79.5	59.7	61.3	33.0	169.3
Percentiles	1	19.5	.3	3.8	14.7	8.7	9.7	1.4	47.9
	5	23.3	1.2	7.9	22.7	16.1	16.3	4.5	59.0
	10	25.9	2.1	9.8	25.5	19.2	19.3	6.5	67.1
	15	27.5	3.3	11.4	27.0	22.0	21.7	7.5	71.0
	20	28.8	4.4	12.4	28.7	24.1	23.5	8.5	75.2
	25	29.5	5.9	13.4	30.5	25.3	25.3	9.0	80.2
	30	30.0	6.7	14.4	31.5	27.0	26.7	10.0	84.5
	35	30.8	7.6	15.4	34.0	28.7	28.3	11.0	88.6
	40	31.8	8.7	16.0	35.5	29.9	29.7	11.5	92.2
	45	32.5	10.2	16.4	37.0	30.7	31.0	12.5	95.7
	50	33.3	11.6	17.4	38.5	32.0	32.2	13.0	99.8
	55	34.0	13.6	18.4	40.0	33.7	33.0	13.9	104.2
	60	35.0	14.9	19.2	42.0	35.0	34.7	14.5	106.8
	65	36.3	15.7	19.9	43.0	36.0	36.0	15.5	109.8
	70	37.0	17.0	21.0	45.0	37.0	37.7	16.0	113.7
	75	38.0	18.6	22.2	46.6	38.0	38.3	17.0	117.2
	80	38.8	20.4	23.2	51.3	39.5	40.0	18.0	121.5
	85	40.0	21.9	24.6	54.0	41.0	42.0	20.5	128.1
	90	41.8	23.6	26.1	58.0	42.7	44.1	21.5	136.8
	95	45.5	25.7	29.9	64.3	49.6	47.6	24.0	147.3
	99	49.4	30.7	37.0	71.8	54.8	57.5	29.2	164.1

Note: The 386 participants are all those who had an error-free trial of the tests.

The collection of data from a number of different recruiting centres in Germany, and the random distribution of test forms provides a representative sample. Controls testing competence remove outliers. These conditions support the descriptive statistics and percentile norms produced in Table 2.4.

Some observations on the distributions

The 386 participants are all those who had an error-free trial of the tests. A very strict criterion of acceptability was applied. If any one of the eight test scores was at a chance level, participants were excluded. These norms apply to those who show complete test-taking competence. One assumes this is a necessary condition of screening procedures. *The Bundeswehr Experimental Test Battery* permits retesting if any one test score is unsatisfactory, by using re-test norms to ensure fairness.

Overall, there is good dispersion, and room at the upper end for discrimination at the top levels. Histogram plots were produced and these are almost completely normal, in all but the most difficult test (*Reasoning Categories*), where there are signs of bottoming.

These norms can be used to produce percentile ranks and automatic reports based on these ranks. Standard scores are a simple function of raw scores transformed to z scores and reformulated to a common mean and variance.

Finally, the combination of adjusted scores and recorded mean latencies for each item type are, in combination, adequate material for adaptive forms of these tests, using a Rasch model or other appropriate metric (Wright, 2002).

Part 3: Validity: Inferences About Function

A number of different forms of validity are needed before the credentials of any test series can be offered as evidence for its effectiveness. Various publications attest to the efficacy of the English language versions (Irvine 2000, a, b; Irvine 2001). Only the report on the paper-and-pencil versions of *The Bundeswehr Experimental Test Battery* exists to show the effects of using these tests with a conscript population (*Arbeitsberichte Psychologischer Dienst der Bundeswehr Nr. 1/2000*).

While the most important form of validity is predictive efficiency, the design of the present study prevents this from being included. A follow-up of those tested would provide the evidence from appropriate training criteria.

However, the following different forms of validity can be assessed here.

Face and Content Validity showing the relevance of the test content to the operational requirements of military service

Construct Validity showing the consistency with which tests are identified with a recognised theoretical entity. Within this broad category there are two subdivisions

Convergent Validity showing the relation of two different types of measures of the same function.

Discriminant Validity showing that the measure is unique in its contribution to what is being tested

Concurrent Validity showing the relationship of new tests to an existing criterion assessed on the same occasion.

Face and Content Validity

Examination of the content of *The Bundeswehr Experimental Test Battery* reveals that it has a low threshold of knowledge and language comprehension, but requires literacy, numeracy and sustained attention to mental procedures. *The Bundeswehr Experimental Test Battery* also provides an extensive exploration of attentional and working-memory skills, because these qualities are the best predictors of learning new materials (Irvine & Christal, 1994; Irvine, 2002). Moreover, conscripts are reported as enjoying the tests themselves.

If one regards the cost-effective screening of motivated conscripts as a sequential operation, these minimum test demands are essential for the completion of basic training. Given biographical data and/or expressed interest in special functions, such as signals, engineering, electronics and computing, for example, tests of comprehension and special knowledge can be reserved for those whose threshold scores justify extra expense in testing and allocation.

Construct Validity

Construct validity largely depends on the structure of the correlations among tests, where there are theoretical markers for the underlying meaning of test scores. For example, the current *Bundeswehr Entry Tests* have well-known markers for General Intelligence, Spatial Ability, Verbal/Educational Ability and Mathematical Principles. Given these, additional tests can be tagged with some certainty.

Similarly, *The Bundeswehr Experimental Test Battery* has a large working memory factor with specific variance in the verbal, numerical and spatial domains.

Table 3.1 Summary Table of Tests and Numbers of Participants Used in Validity Analyses

	Mean	SD	Analysis N	Missing N
EDADJ	32.3	6.9	500	10
ORADJ	16.0	7.1	500	10
RCADJ	10.0	8.6	496	14
NFADJ	36.1	13.8	500	10
OEADJ	28.4	11.9	499	11
WRADJ	28.5	11.8	498	12
TIADJ	11.7	7.0	497	13
ABADJ	91.3	29.4	498	12
VR_THETA Verbal Reasoning Theta	.2	.9	348	162
NUM_THET Number Facility Theta	-.1	1.0	348	162
FR_THETA Figural Reasoning Theta	.3	1.0	348	162
MTR_SCOR Mechanical Comprehension Score	10.7	3.6	348	162
RPR_SCOR Reaction Time Score	51.1	14.1	351	159
RST_SCOR Orthography Test Score	40.6	9.0	218	292
FTR_SCOR Radio Operator Score	60.9	61.1	93	417
SIG_SCOR Signal Detection Score	12.3	5.3	165	345

If the underlying structures of the two batteries are similar, then there is evidence of convergent validity, in that the same construct is being measured by different tests. Where there is uniqueness, or differentiation, then the tests help to discriminate between dimensions, and establish discriminant validity.

In Table 3.1 is a simple summary of the test descriptives for the analyses of structure that follow. While approximately 500 participants were involved in *The Bundeswehr Experimental Test Battery*, complete data from the *Bundeswehr Entry Tests* was available for a maximum of 348 and this reduced to 165 for the *Signal Detection Test* and only 93 for the *Radio Operator Test* Because

of small numbers in other tests, factor analysis was restricted to the tests listed below, using Maximum Likelihood extraction and the objective rotation methods, Promax and Varimax.

Studying the Correlation Matrix

Table 3.2 is used to show the underlying structure and reliability estimates for the item-generative tests for all available participants. The reliabilities are good. The intercorrelations of the tests are very similar to those encountered internationally, and to the paper-and-pencil versions of *The Bundeswehr Experimental Test Battery*. Structurally, the computer-delivered forms reveal a large working memory component, with specific variance associated with verbal, numerical and spatial tasks. Latent structure analyses show a predictable division between working memory tests not dependent upon acquired knowledge, and those tests based upon specific knowledge, such as mechanical and electrical principles tests.

There were also some unexpected findings, in particular quite substantial correlations with age (see Table 3.7). The younger participants are more effective performers than older ones. This could be an artefact of a two year delay in conscripting Russian-speaking migrants from areas of the former USSR (Kazachstan and Siberia) but that is speculation. A variable to record the first spoken language would clarify this.

Correlations with Test Competence are also pronounced (Table 3.6). This constructed variable shows that the underlying domains can be measured by a computer interface outcome. **It is a serendipitous example of trait measurement by more than one method.** It can also be seen as a form of convergent validity – the same thing measured by another method.

Table 3.2 Basic Intercorrelation Matrix for All Core Measures used in Validity Analyses

	EDADJ	ORADJ	RCADJ	NFADJ	OEADJ	WRADJ	TIADJ	ABADJ	VR_THETA Verbal Reasoning Theta	NUM_THET Number Facility Theta	FR_THETA Figural Reasoning Theta	MTR_SCOR Mechanical Comprehension Score	RPR_SCOR Reaction Time Score	RST_SCOR Orthography Test Score	FTR_SCOR Radio Operator Score	SIG_SCOR Signal Detection Score
EDADJ	1.000	.577	.387	.659	.634	.627	.558	.653	.353	.430	.445	.244	.436	.420	.203	.372
ORADJ	.577	1.000	.512	.586	.618	.622	.694	.572	.420	.499	.477	.370	.399	.502	.329	.407
RCADJ	.387	.512	1.000	.428	.442	.468	.567	.458	.445	.442	.468	.333	.360	.442	.216	.243
NFADJ	.659	.586	.428	1.000	.615	.619	.624	.669	.387	.604	.415	.219	.347	.476	.202	.447
OEADJ	.634	.618	.442	.615	1.000	.645	.662	.621	.379	.454	.465	.266	.475	.378	.397	.461
WRADJ	.627	.622	.468	.619	.645	1.000	.658	.588	.475	.484	.459	.324	.413	.520	.315	.336
TIADJ	.558	.694	.567	.624	.662	.658	1.000	.622	.511	.563	.537	.381	.488	.524	.453	.419
ABADJ	.653	.572	.458	.669	.621	.588	.622	1.000	.432	.514	.405	.255	.448	.514	.297	.416
VR_THETA	.353	.420	.445	.387	.379	.475	.511	.432	1.000	.531	.494	.464	.368	.429	.205	.212
NUM_THET	.430	.499	.442	.604	.454	.484	.563	.514	.531	1.000	.588	.389	.346	.528	.444	.289
FR_THETA	.445	.477	.468	.415	.465	.459	.537	.405	.494	.588	1.000	.483	.415	.316	.418	.321
MTR_SCOR	.244	.370	.333	.219	.266	.324	.381	.255	.464	.389	.483	1.000	.295	.221	.315	.218
RPR_SCOR	.436	.399	.360	.347	.475	.413	.488	.448	.368	.346	.415	.295	1.000	.220	.268	.361
RST_SCOR	.420	.502	.442	.476	.378	.520	.524	.514	.429	.528	.316	.221	.220	1.000	.352	.295
FTR_SCOR	.203	.329	.216	.202	.397	.315	.453	.297	.205	.444	.418	.315	.268	.352	1.000	.239
SIG_SCOR	.372	.407	.243	.447	.461	.336	.419	.416	.212	.289	.321	.218	.361	.295	.239	1.000

Construct Validity by Factor Analysis

Already in Table 3.2, the tests were seen to correlate in a consistent fashion. The results of this table can be extended by including the *Bundeswehr Entry Tests* already given by the German Ministry of Defence, and by data reduction methods finding common factors in the complete array of tests used.

Table 3.3 One and Two Factor Solutions for Major Measures

	Factor		Factor	Factor
	1		1	2
TIADJ	.841	NFADJ	.810	.485
WRADJ	.783	OEADJ	.803	.535
ORADJ	.780	ABADJ	.779	.508
OEADJ	.777	EDADJ	.779	.445
ABADJ	.771	TIADJ	.773	.711
NFADJ	.769	WRADJ	.769	.584
EDADJ	.735	ORADJ	.754	.609
RCADJ	.622	RCADJ	.569	.594
NUM_THET Number Facility Theta	.691	RST_SCOR Orthography Test Score	.595	.520
FR_THETA Figural Reasoning Theta	.638	RPR_SCOR Reaction Time Score	.532	.482
RST_SCOR Orthography Test Score	.619	SIG_SCOR Signal Detection Score	.523	.343
VR_THETA Verbal Reasoning Theta	.591	FR_THETA Figural Reasoning Theta	.558	.741
RPR_SCOR Reaction Time Score	.555	NUM_THET Number Facility Theta	.635	.711
SIG_SCOR Signal Detection Score	.506	VR_THETA Verbal Reasoning Theta	.516	.678
FTR_SCOR Radio Operator Score	.448	MTR_SCOR Mechanical Comprehension Score	.340	.633
MTR_SCOR Mechanical Comprehension Score	.441	FTR_SCOR Radio Operator Score	.378	.519

Notes: One Factor Solution Extraction Method: Maximum Likelihood Two factor Solution Extraction Method: Maximum Likelihood. Rotation Method: Oblimin with Kaiser Normalization

Table 3.3 summarises several analyses. First, a word of explanation: not all of the tests used could be analysed together. Several German Ministry of Defence test results had to be omitted because they were not all applied to all recruits or the amount of data collected was insufficient to include them. Next, tests that were composites, such as the *IQ Theta* were omitted from factor analyses because they were dependent on their constituents

The results are all very similar to previous paper-and-pencil analyses from German, USA and UK personnel. There is a plausible case for a single general factor of general capacity to process information

in *The Bundeswehr Experimental Test Battery* and the *Bundeswehr Entry Tests*. Nevertheless, it is possible to extract two factors to provide more clarification of underlying dimensions. One factor identifies with the regular German Ministry of Defence tests and the other with the new tests. The factors correlate at approximately 0.66, indicating the presence of the strong higher-order general factor. This result is very close to the finding realised in the United States. The new tests correlate with the *USA Armed Services Vocational Aptitude Battery* to almost the same degree.

In every study of its tests, The Bundeswehr Experimental Test Battery has identified a robust working memory factor, closely allied to reasoning and fluid intelligence. That interpretation of the first factor is confirmed when the tests are used with a German language population, either in paper-and-pencil or computer-delivered modes.

Working memory is contrasted with the Bundeswehr Entry Tests main factor of intelligence derived from the mental skills and the knowledge expected of those who have completed the secondary school curriculum. The United States *Armed Services Vocational Aptitude Battery* has been shown in previous extensive studies to produce a verbal-educational construct that is very much a function of acquired knowledge. Whether or not this interpretation may be applied to the originally constructed German Ministry of Defence tests in the second factor is an academic matter. For all practical purposes, the second factor identifies elements of a factor dependent on skills in the German language.

In summary, Table 3.3 reproduces the now familiar structure for *The Bundeswehr Experimental Test Battery* when compared with conventional tests of aptitude. The two basic factors are reproduced. The first factor helps to provide convergent validity for two of the *Bundeswehr Entry Tests*. This is the underlying Working Memory factor of *The Bundeswehr Experimental Test Battery* which is called *Capacity to Process Information*.

The Signal Detection test (Sig_T) of the Bundeswehr Entry Tests is given firmer definition as a form of working memory test, as is the Choice Reaction Time test (RP_S). Both of these have higher loadings on *The Bundeswehr Experimental Test Battery* working memory factor than the verbal aptitude factor underlying many of the *Bundeswehr Entry Tests*.

Tests that load equally well on both factors include the Transitive Inference and the Orientation Test . The Mathematical Principles test of the Bundeswehr Entry Tests is similarly a pivotal measure with loadings on both factors.

The two constructs correlate at .66 to .70 in this analysis, a result that is consistent with other studies carried out in the USA and in the United Kingdom. The extent of this correlation indicates the presence of a higher-order general factor of intelligence commonly found in populations of extended range. Conscripts provide such a population.

To conclude, there is convergent validity because different tests and approaches measure the same dimensions. These are also signs of discriminant validity or uniqueness in the Bundeswehr Entry Tests of Signal Detection and Choice Reaction Times. This uniqueness was identified by the presence of the working memory markers of The Bundeswehr Experimental Test Battery.

Concurrent Validity by Regression

Operational definitions of what may be measured by the new tests can be gathered from attempts to predict scores on tests used by the German Ministry of Defence. The new tests are used as independent variables, and the indices provided by the tests actually in use by the German Ministry of Defence as dependent variables.

The *Intelligence Theta* score can be predicted from the results of four tests with a multiple R of .70. This is close to the value for paper-and-pencil tests in the previous study (Irvine, Kutschke & Walker, 2000). Because a stepwise regression function was fitted, relatively small sample numbers ensure that the criterion for inclusion is reached quickly.

Table 3.4 is a simple summary of regression analyses performed on all German Ministry of Defence measures included, regardless of sample size. Each coefficient is a multiple R found by regressing the new tests against the score for the target test. Table 3.4 shows what components of *The Bundeswehr Experimental Test Battery* are the best predictors of the *Bundeswehr Entry Test* scores and the order in which they entered the equation. Because the sample sizes vary, the results should be interpreted with caution. They are nevertheless confirmed as useful values, being of the same order as those found in the paper-and-pencil study (Irvine et al., 2000).

What proved to be the most difficult test, *Reasoning Categories*, was excluded from these analyses. When it was used out of curiosity, it assumed the functions of the *Transitive Inference* or *Deductive Reasoning Test*, which was designed as a suitable replacement. *Reasoning Categories* and *Transitive Inference* are equally powerful tests when administered at the correct level.

These coefficients reveal the power of the new item-generative tests. They also show that some of the tests used by the German Ministry of Defence have specific variance that is not captured by the new tests. These tests include *Mechanical and Electrical Principles*, *Auditory Thresholds* and so on.

Nevertheless, the item-generative tests can readily be substituted for screens for intelligence or information processing capacity. Such a decision could be taken with some confidence in the outcomes for accuracy, fairness and low cost of production; and of security through regular renewal by simple insertion of parallel forms without restandardisation.

Moreover, the overall cost-benefits of using only a few tests of short duration as a mass screening strategy are considerable. **For example, to reach a correlation of .70 with INT_THETA took only 19 minutes of testing time. This translates to a test session of 30 minutes after instructions and practice items.** Extended assessment can then be reserved for those conscripts eligible for the much smaller number of specialist military roles.

Table 3.4: Regression of The Bundeswehr Experimental Test Battery upon key Components of Bundeswehr Entry Tests

Regressions	INT_THET Intelligence		VR_THETA Verbal Reasoning		NUM_THET Number Facility		FR_THETA Figural Reasoning		RST_SCOR Orthography Test Score		MTR_SCOR Mechanical Comprehension Score		SIG_SCOR Signal Detection Score		RPR_SCOR Reaction Time Score	
R	.70		.56		.65		.57		.63		.41		.52		.54	
R ²	.48		.32		.42		.32		.39		.17		.26		.29	
Tests and Order In Equation	TIADJ	NFADJ	TIADJ	WRADJ	NFADJ	TIADJ	TIADJ	EDADJ	TIADJ	ABADJ	TIADJ	ORADJ	OEADJ	EDADJ	TIADJ	OEADJ
	WRADJ	ORADJ	ABAADJ				ORADJ	WRADJ						EDADJ		

Table 3.5: Effects of Education Level on The Bundeswehr Experimental Test Battery and Bundeswehr Entry Tests Scores

Education		EDADJ	ORADJ	RCADJ	NFADJ	OEADJ	WRA DJ	TIADJ	ABA DJ	INT_THET Intelligence	VR_THETA Verbal	NUM_THET Number	FR_THETA Figural	MTR_SCOR Mechanical	RST_SCOR Orthography
1.0	Mean	29.2	12.3	5.5	28.9	23.6	21.5	8.1	75.0	-1.6	-.5	-.8	-.3	9.3	31.7
	N	99.0	99.0	98.0	99.0	99.0	99.0	99.0	99.0	79.0	80.0	80.0	80.0	80.0	25.0
	SD	6.6	6.4	6.0	11.3	12.5	11.2	5.2	24.9	1.7	.8	.7	.7	3.1	8.9
2.0	Mean	31.4	15.3	9.8	35.2	27.6	27.8	11.2	88.3	.1	.2	-.3	.2	10.2	38.4
	N	172.0	172.0	171.0	172.0	171.0	172.0	171.0	172.0	160.0	160.0	160.0	160.0	160.0	102.0
	SD	6.4	6.7	8.2	12.9	11.4	10.6	6.3	26.9	1.7	.7	.6	.9	3.4	8.2
3.0	Mean	36.4	20.4	14.3	44.3	34.2	35.5	16.2	111.2	2.3	.8	.6	.9	12.4	45.6
	N	112.0	112.0	112.0	112.0	112.0	112.0	112.0	112.0	108.0	108.0	108.0	108.0	108.0	91.0
	SD	6.6	6.8	8.9	13.1	10.1	9.7	6.0	27.2	2.0	.7	1.0	1.0	3.5	6.9

Total	Mean	32.3	16.0	10.0	36.2	28.5	28.4	11.8	91.6	.4	.2	-.1	.3	10.7	40.6
	N	383.	383.	381.	383.	382.	383.	382.	383.	347.0	348.0	348.0	348.0	348.0	218.0
		0	0	0	0	0	0	0	0						
	SD	7.1	7.3	8.6	13.8	12.0	11.7	6.7	29.7	2.3	.9	1.0	1.0	3.6	9.0
	Eta	.355	.376	.338	.395	.252	.416	.420	.450	.573	.457	.552	.392	.297	.543
	Eta Squ.	.126	.142	.115	.156	.063	.173	.176	.202	.328	.209	.304	.153	.088	.295

Table 3.6: Effects of Test Competency on Scores for *The Bundeswehr Experimental Test Battery* and *Bundeswehr Entry Tests*.

Test Competence		EDAD J	ORAD J	RCAD J	NFAD J	OEAD J	WRA DJ	TIADJ	ABA DJ	INT_THET Intelligence	VR_THETA Verbal Reasoning	NUM_THET Number Facility	FR_THETA Figural Reasoning	MTR_SCOR Mechanical Comprehension Score	RST_SCOR Orthograph y Test Score
1 Satisfactory	Mean	33.6	18.0	12.6	39.6	31.9	31.9	13.6	100.	1.0	.4	.1	.5	11.2	41.8
	N	354.	354.	354.	354.	354.	354.	354.	354.	235.0	236.0	236.0	236.0	236.0	169.0
	SD	6.5	6.6	8.1	12.9	9.7	9.9	6.1	26.8	2.1	.8	.9	.9	3.6	8.9
2 One or More Neg.	Mean	29.1	11.2	3.8	27.7	19.9	20.2	7.0	69.8	-1.0	-.2	-.6	-.1	9.5	36.6
	N	146.	146.	142.	146.	145.	144.	143.	144.	112.0	112.0	112.0	112.0	112.0	49.0
	SD	7.0	6.1	6.3	12.3	12.5	12.1	6.8	24.1	2.2	.8	.9	.9	3.1	8.3
Total	Mean	32.3	16.0	10.0	36.1	28.4	28.5	11.7	91.3	.4	.2	-.1	.3	10.7	40.6
	N	500.	500.	496.	500.	499.	498.	497.	498.	347.0	348.0	348.0	348.0	348.0	218.0
	SD	6.9	7.1	8.6	13.8	11.9	11.8	7.0	29.4	2.3	.9	1.0	1.0	3.6	9.0
	Eta	.297	.439	.462	.391	.458	.450	.434	.466	.404	.353	.336	.311	.224	.240
	Eta Sq.	.088	.193	.214	.153	.210	.202	.188	.218	.163	.125	.113	.097	.050	.057

Table 3.7: Age Group Means and Effect sizes for *The Bundeswehr Experimental Test Battery* and *Bundeswehr Entry Tests*

AGEG RP Age Group		INT_THET Intelligen ce Theta	VR_THETA Verbal Reasoning Theta	NUM_TH ET Math. Theta	FR_THETA Figural Reas. Theta	MTR_SC OR Mech Comp. Score	RST_SC OR Orthog .Test Score	SIG_SC OR Signal Detecti on Score	EDADJ	ORADJ	RCADJ	NFADJ	OEADJ	WRADJ	TIADJ	ABADJ
1.0 To 17	Mean	1.3	.5	.3	.5	11.1	44.2	12.8	34.2	18.3	12.1	39.6	31.2	31.6	14.1	100.9
	N	102.0	103.0	103.0	103.0	103.0	78.0	58.0	171.0	171.0	170.0	171.0	171.0	170.0	170.0	170.0
	SD	2.5	.8	1.1	1.0	3.6	8.3	5.2	7.5	7.0	8.8	15.1	12.4	12.2	6.7	29.7
2.0 18yrs	Mean	.5	.2	-.2	.4	10.9	38.9	13.0	32.2	15.2	9.9	36.2	28.0	28.7	11.7	91.7
	N	80.0	80.0	80.0	80.0	80.0	52.0	39.0	110.0	110.0	110.0	110.0	110.0	110.0	110.0	110.0
	SD	2.0	.8	.8	.9	3.4	9.3	5.3	6.0	7.5	8.4	13.9	11.5	11.9	6.9	30.1
3.0 19yrs	Mean	-.2	.1	-.4	.2	10.5	37.4	11.4	30.9	15.1	8.7	32.9	27.0	26.2	9.9	84.4
	N	101.0	101.0	101.0	101.0	101.0	56.0	46.0	133.0	133.0	131.0	133.0	132.0	132.0	131.0	132.0
	SD	2.0	.9	.7	.8	3.6	8.0	5.2	6.7	6.8	8.2	11.9	11.1	10.6	7.0	26.3
4.0 20+yr s.	Mean	-.8	-.1	-.6	-.1	9.8	39.3	11.2	30.9	13.6	8.1	33.4	25.2	25.2	9.6	82.0
	N	54.0	54.0	54.0	54.0	54.0	26.0	19.0	80.0	80.0	79.0	80.0	80.0	80.0	80.0	80.0
	SD	2.1	.9	.9	.9	3.4	8.6	5.8	6.5	6.2	8.2	11.9	11.6	11.1	6.1	27.5
Total	Mean	.3	.2	-.2	.3	10.6	40.5	12.2	32.3	16.0	10.1	36.0	28.4	28.5	11.7	91.4
	N	337.0	338.0	338.0	338.0	338.0	212.0	162.0	494.0	494.0	490.0	494.0	493.0	492.0	491.0	492.0
	SD	2.3	.9	.9	1.0	3.5	9.0	5.3	6.9	7.1	8.6	13.8	11.9	11.8	7.0	29.5
	R	-.336	-.245	-.340	-.233	-.121	-.264	-.126	-.195	-.235	-.180	-.194	-.184	-.214	-.261	-.253
	R Sq.	.113	.060	.116	.054	.015	.070	.016	.038	.055	.032	.038	.034	.046	.068	.064
	Eta	.337	.247	.349	.235	.126	.324	.146	.207	.251	.185	.208	.188	.218	.271	.259

Eta Sq.	.113	.061	.122	.055	.016	.105	.021	.043	.063	.034	.043	.035	.047	.073	.067
----------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------

Concurrent Validity by Group Classification

Because subjects do not come to test situations with nothing in their heads (or so one hopes), validity can be estimated by group membership when this membership is what we have described elsewhere as a low-inference variable.

Education Level

Level of education was recorded for several participants. Three finite groups were identified: Hauptschulabschluss (Group1); Realschulabschluss (Group 2); and (Allgemeine Hochschulreife Group3 including University/College Entry Qualifications).

The outcomes of ANOVAs for the effect of level of education on all of *The Bundeswehr Experimental Test Battery* and on the *Bundeswehr Entry Tests* (Intelligence, Verbal Reasoning, Mathematical Facility, Figural Reasoning Mechanical Comprehension and Orthography) are shown in Table 3.5. Predictably, every F ratio, with one exception, was highly significant; and effect sizes separating top and bottom groups were large.

Test Competence

An unusual group classification on this occasion was the filter variable Test Competence. Two groups were created: those who had no chance score or equivalent in any of The Bundeswehr Experimental Test Battery tests and those who had one or more scores failing to meet the criterion of acceptability. The inference is that any failure to produce a score within acceptable boundaries is indicative of a larger set of skill deficiencies.

Table 3.6 reveals substantial effect sizes separating the two groups. These are confirmed by the strength of the Eta correlations between test competence and performance on *The Bundeswehr Experimental Test Battery* and the *Bundeswehr Entry Tests*.

Age at Conscriptio

Additionally, the consistent negative correlations with age (Table3.7) are unusual and one can only suppose that age is a package variable for some underlying social factor. Older conscripts have among their number immigrants whose first language is not German. That being so, age categories are additional evidence of test validity by language group membership.

Validity by Report Profiles

Extensive testing with an applicant sample of over 2000 USA servicemen subjects produced stable norms for *Error Detection*, *Orientation*, *Reasoning Category*, *Number Fluency*, *Word Rules*, *Odds and Evens* and *Transitive Inference*. These norms were used to produce a computer-based report system. The individual reports provide both summary and detailed commentary on test performance. The reports can be used to guide allocation decisions.

Group Summaries

As an illustration, a group of 18 participants was chosen at random and detailed study of their results was undertaken. Here the selected group of 18 is ranked by *Capacity to Process Information* score (CPI). This and the individual test scores are listed in the extract above from the actual report system. The system permits automatic ranking by CPI of all participants on the database.

Table 3.8 Groups by Bundeswehr Entry and The Bundeswehr Experimental Test Battery Scores

Ability Group		N	Minimum	Maximum	Mean	Std. Dev.
3.0	INTTHETA IQ BET	6.0	.7	6.1	2.9	1.9
	PCTILE Percentile Rank	6.0	80.0	90.0	84.2	3.8
	CPI_STD CPI Standard Score	6.0	113.3	125.0	117.3	4.3
2.0	INTTHETA IQ BET	6.0	-1.0	3.8	2.3	1.9
	PCTILE Percentile Rank	6.0	50.0	80.0	61.7	10.3
	CPI_STD CPI Standard Score	6.0	100.4	112.6	105.1	4.1
1.0	INTTHETA IQ BET	5.0	-1.7	1.0	-.7	1.1
	PCTILE Percentile Rank	5.0	1.0	25.0	13.2	10.1
	CPI_STD CPI Standard Score	5.0	75.9	89.6	83.9	6.3

The *Intelligence Theta* scores for the top, middle and low groups of participants were collated from records provided by members of the team. Table 3.8 shows the results of grouping these participants by *Capacity to Process Information* standard scores into top, middle and bottom thirds, with approximately 6 in each group. There are clear demarcations among the groups and the summary reflects these classifications. The results show large effect sizes of nearly two standard deviations of INT_THETA scores. The clear-cut divisions among the three groups

reflect the positive correlations between *Bundeswehr Entry Tests* and *The Bundeswehr Experimental Test Battery*.

Individual Reports

Quantitative divisions are also reflected in the quality of the reports produced for individual members of these groups. Differences among the reports for members of the different ability groups provided further evidence of report validity.

Part 4: Summary and Conclusions

The item-generation principles demonstrated here are no longer experimental. A comprehensive publication on the principles and applications of Item-generation theory is now available (Irvine and Kyllonen, 2002). It shows that item-generative methods are in use by a number of independent agencies.

In technical terms, the results of The Bundeswehr Experimental Test Battery can be summarised as follows

- The tests can be generated in German; and they behave just as the English versions do in similar delivery contexts.
- Their attributes include high reliabilities, consistent construct validities, convincing concurrent validities, sensitivity to group membership variables, and coherent reports of subjects chosen by the German Ministry of Defence Standard 7-point intelligence criterion.
- The reliabilities of individual tests range from very good to excellent.
- The content, construct and concurrent validities are consistent with previous Bundeswehr paper-and-pencil trials.
- In conjunction with computer-generated scoring and storage mechanisms, the tests coherently grade conscripts and volunteers eligible to serve in the German Armed Services.
- The report system provides valid diagnostic information to permit counselling and placement in appropriate training contexts.

Practical and social implications may now be considered.

Item-Generative tests ensure security at little or no administrative cost. Adaptive tests used without automatic item generation in computer-delivered contexts have major security difficulties and they are costly to maintain to prevent compromise (Wainer, 2002). More important than adaptive testing in computer-delivery is security against theft, compromise or cheating. The use of item-generative theory to produce multiple parallel forms (Irvine, 2002), provides a guarantee against destruction of tests by illegal and/or unethical means.

Because the tests resist compromise, there are additional social benefits. Fairness is supported by the use of retest norms on second attempts. In addition, through the availability of pre-test materials, test demands can be open and transparent. Pre-test information can be distributed ahead of time because no one will know the answers to the test items produced on the day, not even the test administrators. Finally, in *The Bundeswehr Experimental Test Battery* the scores of persons tested are not limited by lack of knowledge, only by the capacity to process new information in a short time span.

References and Bibliography

- Arbeitsberichte Psychologischer Dienst der Bundeswehr Nr. 1/2000. See Irvine, Kutschke and Walker, below.
- Armstrong, R. D., Jones, D.H., & Wang, Z. (1994). Automated parallel test construction using classical test theory. *Journal of Educational Statistics*, **19**: 73-90.
- Baddeley, A.D. (1968). A three-minute reasoning test based on grammatical transformation. *Psychonomic Science*, **10**: 341-342.
- Baddeley, A.D. & Hitch, G.(1974). Working memory. In G.H. Bower, (Ed.), *The Psychology of learning and motivation*, Vol.8, 47-90.
- Bartram, D. (1987). The development of an automated testing system for pilot selection: the MICROPAT project. *Applied Psychology: International Review*: 36; 279-298.
- Bejar, I.J. (1986a). *The psychometrics of mental rotation: (RR-86-19)*. Educational Testing Service: Princeton, N.J.
- Bejar, I.J. (1986b). *Analysis and generation of Hidden Figure items: A cognitive approach to Psychometric Modelling. (RR-86-20)*. Educational Testing Service: Princeton, N.J.
- Bejar, I.J. (1986c). *Final Report: Adaptive testing of spatial abilities. (ONR 150 531)*. Educational Testing Service: Princeton, N.J.
- Bejar, I.; & Yocom, P. (1991). A generative approach to the modelling of isomorphic hidden-figure items. : *Applied Psychological Measurement*; 1991 Jun Vol **15**(2) 129-137
- Bevans, H.G. (1966). *Probability (Confidence) scoring for the Standard Progressive matrices and the Advanced Matrices*. Paper presented to the British Psychological Society Annual Conference, Swansea, Wales.
- Bongers, S.H. & Greig, J.E (1997). *An Australian trial of the British Army Recruit Battery – Part 2*. Report of The Air Force Office, Mawson; ACT 2607, Australia
- Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). *The four generations of computerised educational measurement* . Research Report (RR88-35) Educational Testing Service, Princeton, NJ.
- Carroll, J.B. (1976). Psychometric tests as cognitive tasks: a new "Structure of Intellect." In L.B. Resnick, (Ed.), *The nature of intelligence*. Hillsdale, NJ: Erlbaum.
- Carroll, J.B. (1980). *Individual difference relations in psychometric and experimental cognitive tasks*. Report No. 163, Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, N.C.27514.
- Carroll, J.B. (1983). The difficulty of a test and its factor composition revisited. In H. Wainer & S. Messick, (Eds.), *Principals of modern psychological measurement*. Hillsdale, NJ: Erlbaum.
- Carroll, J.B. (1986). Defining abilities through the person characteristic function. In S.E. Newstead, S.H. Irvine & P.L. Dann, (Eds.). *Human Assessment: Cognition and motivation*. Dordrecht, Netherlands: Nijhoff.
- Carroll, J.B. (1987). New perspectives in the analysis of abilities. In R.R. Ronning, J.A. Glover, J.C. Conoley, & J.C. Witt (Eds.), *The influence of cognitive psychology on testing*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Carroll, J.B. (1993). *Human cognitive abilities: a survey of factor-analytic studies*. Cambridge, New York.
- Carroll, J.B., Meade, A., & Johnson, E.S. (1991). Test analysis with the person characteristic function: implications for defining abilities. In R.E. Snow & D.E. Wiley (Eds.), *Improving inquiry in education, psychology and social science: a book in honour of Lee J. Cronbach* (pp.109-143).Hillsdale, NJ: Erlbaum

- Christal, R.E. (1984). *New cognitive tests being evaluated by TTCP services*. Report to the Technical Cooperation Program Meeting of 1984, Armstrong Laboratory, Brooks AFB, San Antonio, Texas.
- Christal, R.E. (1987). *A factor-analytic study of tests of working memory*. Unpublished Report; Human Resources Division, USAF Armstrong Laboratory, Brooks AFB, San Antonio, Texas.
- Christal, R.E. (1990). *Comparative validities of ASVAB and LAMP tests for Logic Gates learning*. Technical Report, Armstrong Laboratory, Brooks AFB San Antonio, Texas.
- Clark, H.H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, **76**: 387-404.
- Clark, H.H. (1970). Comprehending comparatives. In: G.B Flores D'Arcais & W.J.M. Levelt (Eds.), *Advances in Psycholinguistics*. Amsterdam: North Holland.
- Clark, H.H. & Chase, W.G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, **3**: 472-517.
- Collis, J.M. & Irvine, S.H. (1991a). *Predictive Validity and Utility of the ABC Battery with Royal Navy Officers under Training*. SP(N) Report TR 261 - 1991
- Collis, J.M. & Irvine, S.H. (1991b). *The Plymouth ABC Battery for Artificer Apprentice Entrants*. Validity and reliability studies. SP(N) Report TR 265 - 1991
- Collis, J.M. & Irvine, S.H. (1991c). *The Plymouth ABC Battery for RN Non-Technician Ratings under Training*. Validity and reliability studies. SP(N) Report TR 266 - 1991
- Collis, J.M. & Irvine, S.H. (1991d). *The Plymouth ABC Battery for WRNS Rating Entrants*. Validity and reliability studies. SP(N) Report TR 267 - 1991.
- Collis, J.M. & Irvine, S.H. (1991e). *The Plymouth ABC Battery for RN/WRNS Ratings and RM other Ranks under Training*. Validity and reliability studies: summary report SP(N) Report TR 271 - 1991
- Collis, J.M. & Irvine, S.H. (1993a). *The ABC Combined Battery for Artificer Apprentices*. Further validity studies. SP(N) Report TR 311 - 1993
- Collis, J.M. & Irvine, S.H. (1993b). *The ABC Combined Battery for Royal Marine Apprentices*. Further validity studies. SP(N) Report TR 312 - 1993
- Collis, J.M. & Irvine, S.H. (1993c). *The ABC Combined Battery for RN/WRNS Non-Technicians*. Further validity studies. SP(N) Report TR 313 - 1993
- Collis J.M. & S. H. Irvine. (1993d). *The Effects of Pre-Knowledge, Retest and Types of Test Administration on Computer Generated Cognitive Tasks in a Group of Royal Navy and Royal Marine Entrants*. SP(N) Report TR 314
- Collis, J.M. & Irvine, S.H. (1994). *A New Generation of Ability Tests for Selection and Training*. The Navy Personnel Series. HAL Technical Report 1:1994.
- Collis, J. M., Tapsfield, P.G.C., Irvine, S.H., Dann, P.L. & Wright, D. (1995). The British Army Recruit Battery goes operational: from theory to practice in computer-based testing using item-generation techniques. *International Journal of Selection and Assessment*, **3**: 96-103.
- Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist*, **12**: 671-684.
- Dann, P.L. & Irvine, S.H. (1986). *Handbook of Computer-based Cognitive Tasks*. Centre for Computer-Based Assessment, University of Plymouth, Plymouth, Devon, UK.
- Dennis, I. (1993). *The Development of an Item Generative Test of Spatial Orientation Closely Related to Test SP80.A*. SP(N) Report TR 307
- Dennis, I. (1995). *The structure and development of numeracy and literacy tests in the Navy Personnel Series*. Human Assessment Laboratory, University of Plymouth.
- Dennis, I., Collis, J. M., & Dann, P.L. (1995). Extending the scope of item generation to tests of educational attainment. *Proceedings of the International Military Testing Association*, Toronto.
- Dennis, I. & Evans J.StB.T (1989). *System architecture for computerised assessment*. Human Assessment Laboratory, University of Plymouth Report for The Army Personnel Research Establishment (Contract 2021/12). Plymouth, UK.
- Dennis, I. & Evans, J. St.B. T. (1996). The speed-error trade off problem in psychometric testing. *British Journal of Psychology*, **87**: 105-129.
- Dennis I., & Tapsfield, P.G.C. (Eds.) (1996). *Human abilities, their nature & measurement*. Erlbaum, Hillsdale, NJ

- Embretson, S.E. (1995). Working memory capacity versus general control processes in abstract reasoning. *Intelligence*, **20**: 169-189.
- Embretson, S.E. (1996). Multidimensional latent trait models in measuring fundamental aspects of intelligence. In Dennis I., & Tapsfield, P.G.C. (Eds.) *Human abilities, their nature & measurement*. Erlbaum, Hillsdale, NJ.
- Evans, J.St.B.T. (1982). *The psychology of deductive reasoning*. London: Routledge.
- Evans, J. St.B.T. & Wright, D.E. (1992). *The Transitive Inference Task*. HAL Technical Report 2-1992 (APRE).
- Evans, J. St.B.T. & Wright, D.E. (1993). *The Properties of Fixed-Time Tests: A Simulation Study*. HAL Technical Report (APRE).
- Eysenck, H.J. (Ed.). (1982). *A model for intelligence*. New York: Springer-Verlag.
- Furneaux, W.D. (1952). Some speed error and difficulty relationships within a problem-solving situation. *Nature*, **170**: 3.
- Goeters, K-M, (1979). *Die Aenderung Der Psychometrischen Kennwerte Und Der Faktorenstruktur Als Folge Der Uebung Von Tests*. PhD Dissertation, University of Hamburg, Hamburg,
- Goeters, K-M, & Rathje, H. (1992). *Computer-Generierte Parallel-Tests Fur Die Faehigkeitsmessung In Der Eignungsauswahl Von Operationellem Luftfahrtpersonal*. DLR Institut fur Flugmedizin Abteilung Luft-und Raumfahrtpsychologie, Hamburg.
- Greig, J.E. & Bongers, S.H. (1996). An Australian trial of the British Army Recruit Battery. *Proceedings of the 38th Annual Conference of the International Military Testing Association*. San Antonio, November 1996.
- Grenzebach, A. P.; & McDonald, J. E. : (1992). Alphabetic sequence decisions for letter pairs with separations of one to three letters. *Journal of Experimental Psychology Learning, Memory, and Cognition*, Vol **18**(4) 865-872
- Groen, G.J., & Parkman. J.M. (1972). A chronometric analysis of simple addition. *Psychological Review*, **79**: 329-343.
- Harris, R.L. & Tapsfield, P. G. C. (1995). *The British Army Recruit Battery Trials of Pre-Test Booklets*. Human Assessment Laboratory, University of Plymouth Technical Report 10-1995. Plymouth, UK.
- Holroyd, S. R., Atherton, R.M. & Wright, D.E. (1995a). The criterion related validity of the British Army Recruit Battery. *Proceedings of the 34th Annual Conference of the International Military Testing Association*. Toronto, October 1995.
- Holroyd, S. R., Atherton, R.M. & Wright, D.E. (1995b). Validation of the British Army Recruit Battery against measures of performance in basic military training. Centre for Human Sciences, Report DRAJCHS/liS3/CR95019/1.0. DRA, Famborough.
- Hockey, G.R.J. & Maclean, A. (1986). Direct temporal analysis of individual differences in cognitive skill. In S.E. Newstead, S.H. Irvine & P.L. Dann, (Eds.), *Human assessment: cognition & motivation*. Dordrecht: Nijhoff.
- Hornke, L.F. & Habon, M.W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*, **10**, 369-380.
- Hockey, G.R.J., Maclean, A. & Hamilton, P. (1981). State changes and the temporal patterning of component resources. In J. Long & A.D. Baddeley, (Eds.), *Attention and performance, Vol. 9*. Hillsdale, NJ: Erlbaum.
- Hough, P.V.C. (1962). *Method and means for recognising complex patterns*. U.S. Patent 3,069,654.
- Hunt, E., Lunneborg, C. & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology*, **7**: 194-227.
- Irvine, C.D. & Irvine, S.H. (1996). Effects of antihypertensive treatment on cognitive function of older patients: effect is not proved. *British Medical Journal*, **313**: 166.
- Irvine, S.H. (1998a). *The computer-generation of ability tests for adaptive testing in selection and training: A report in the form of a technical handbook*. USAF Air Force Laboratory, Brooks AFB, San Antonio, Texas 78235.
- Irvine, S.H. (1998b). *New tests for recruitment: standardisation and validation*. Final Report for The Royal Ulster Constabulary. Inpsych Ltd. Dawlish, Devon, UK.
- Irvine, S.H. (2002). The foundations of item generation for mass screening. In Irvine, S.H. & Kyllonen, P.C. (Eds.), (2002). *Item generation for test development* Erlbaum Associates, Mahwah, NJ pp.3-34.
- Irvine, S.H. & Berry. J.W., (Eds.), (1988b). *Human abilities in cultural context*. New York:Cambridge.
- Irvine, S.H. & Christal, R.E. (1994). *The Primacy of Working Memory in Learning to Identify Electronic Logic Gates*. HAL Technical Report 4 1994-95. University of Plymouth.

- 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.
- Irvine, S.H., Dann, P.L. & Evans, J. St.B. T. (1987). *Item Generative Approaches for Computer-Based Testing: A Prospectus for Research*. Report for the Army Personnel Research Establishment, Plymouth Polytechnic.
- Irvine, S.H., Dann, P.L., Evans, J. St.B.T., Dennis, I., Collis, J., Thacker, C. & Anderson, J.D. (1989). *Another Generation of Personnel Selection Tests. Stages in a new theory of computer-based test construction*, Human Assessment Laboratory, University of Plymouth, Plymouth, Devon.
- Irvine, S.H., Kutschke, T. & Walker, R.F. (2000). *Screening Conscripts in Germany Using Item-generative Tests*. (Arbeitsberichte Psychologischer Dienst der Bundeswehr Nr. 1/2000). Bundesministerium der Verteidigung, Bonn, Deutschland.
- Irvine, S.H. & Kyllonen, P.C. (Eds.), (2002). *Item generation for test development* Erlbaum Associates, Mahwah, NJ.
- Irvine, S.H. & Newstead, S.E. (Eds.), (1987b). *Intelligence and cognition: contemporary frames of reference* Dordrecht, Netherlands: Nijhoff.
- Irvine, S.H. & Reuning, H. (1981). Perceptual speed and cognitive controls. *Journal of Cross-Cultural Psychology*, **12**, 425-444.
- Irvine, S.H., Schoeman, A. & Prinsloo, W. (1988). Putting cognitive theory to the test: group testing reassessed using the cross-cultural method. In G.K. Verma & C. Bagley (Eds.). *Cross-cultural studies of personality, attitudes and cognition*. Macmillan: London.
- Irvine, S.H., Wulf, J., Schambach, S., Kutschke, T., & Walker, R.F. (2003). *Screening conscripts in Germany using multiple forms of item-generative computer-delivered tests*. Progress report to Bundesministerium der Verteidigung, Bonn, Deutschland.
- 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/fiS3/CR96049/1.0 Defence Evaluation and Research Agency, Farnborough.
- Jacobs, N.R., Cape, L.T. and Lawton, D.H. (1997). Validation of the British Army Recruit Battery (BARB) against phase 2 military training performance measures. Centre for Human Sciences, Report PLSD/CHS/HS3/CR97018/1.0. Defence Evaluation and Research Agency, Farnborough.
- Jensen, A.R. (1982). Reaction time and psychometric g. In H.J. Eysenck, (Ed.), *A model for intelligence*. Springer-Verlag: New York.
- Jensen, A.R. (1988). Speed of information-processing and population differences. In Irvine, S.H. & Berry, J.W, (Eds.). *Human abilities in cultural context*. New York: Cambridge.
- Just, M.A. & Carpenter, P.A. (1985). Cognitive co-ordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, **92**: 137-172.
- Kirsch, H. (1971). Der Wegfiguren Test als Auswahlinstrument. *Aviation Psychology*. (Quoted in Goeters, K_M. 1979, p.125).
- 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/ HS3/CR96060/1.0. DRA, Farnborough.
- Kornbrot, D. E. (1988). Random walk models of binary choice: the effect of deadlines in the presence of asymmetric payoffs. *Acta Psychologica*, **69**: 109-127.
- Kornbrot, D. E. (1989). Organisation of keying skills: the effect of motor complexity and number of units. *Acta Psychologica*, **70**: 19-41.
- Kornbrot, D. E. (1997). Information accrual models of cognitive processes: evidence from the shape of reaction-time distributions. Manuscript for publication, Department of Psychology, University of Hertfordshire, United Kingdom.
- Kyllonen, P.C. (1986). Theory-based cognitive assessment. In J. Zeidner (Ed.), *Human productivity enhancement: Organisations, personnel & decision-making, Vol.1*. New York: Praeger.
- Kyllonen, P. C. (1996). Is working memory capacity Spearman's g? In Dennis I., & Tapsfield, P.G.C. (Eds.). *Human abilities, their nature & measurement*. Erlbaum, Hillsdale, NJ.
- Kyllonen, P. C. & Christal, R.E. (1988). *Cognitive modelling of learning abilities: A status report of LAMP. (AFHRL-TP-87-66)*. Brooks AFB, Texas: Manpower & Personnel Division, Air Force Human Resources Laboratory.

- Kyllonen, P.C. & Christal, C.E. (1989). Cognitive modelling of learning abilities. In R. Dillon & J.W. Pellegrino, (Eds.). *Testing: Theoretical and applied issues*, San Francisco, Freeman.
- Kyllonen, P.C. & Woltz, D.J. (1988). *Role of cognitive factors in the acquisition of cognitive skill*. Paper delivered at Minnesota Symposium on Learning and Individual Differences: University of Minnesota, Minneapolis, April 14-16.
- Lohman, D. F. (1994). Component scores as residual variation; (or why the intercept correlates best). *Intelligence*, **19**: 1-11.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Miller, G. A., & McKean, K.E. (1964). A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology*, **16**: 297-308.
- Mislevy, R.J. & Sheehan, K.M. (1988). *The role of collateral information about examinees in item parameter estimation*. ETS Research Report (RR-88-55-ONR). Educational Testing Service, Princeton, NJ.
- Mislevy, R.J. & Wingersky, M.S., Irvine, S.H. & Dann, P.L. Resolving mixtures of strategies in spatial visualisation tasks. *British Journal of Mathematical and Statistical Psychology*, **44**, 265-288.
- Moyer, R.S. & Landauer, T.K. (1967). Time required for judgements of numerical inequality. *Nature*, **215**: 1519-1520.
- Neimark, E.D. & Estes, W.K. (1967). *Stimulus sampling theory*. Holder; San Francisco.
- Parkman, J.M. (1972). Temporal aspects of simple multiplication and comparison. *Journal of Experimental Psychology*, **95**: 437-444.
- Posner, M.L., Boies, S.J., Eichelman, W.H., & Taylor, R.J. (1969). Retention of visual name codes of single letters. *Journal of Experimental Psychology Monographs*, **79**: 1 (Part2) 1-16.
- Prince, M.J. , Bird, A.S., Blizzard, R.A., & Mann, A. H. (1996). Is the cognitive function of older patients affected by hypertensive treatment? *British Medical Journal*, **312**: 801-808..
- Resnick, L.B. (Ed.). (1976). *The nature of intelligence*. Hillsdale, NJ: Erlbaum.
- Restle, F. & Davis, J. H. (1962). Success and speed of problem-solving by individuals and groups. *Psychological Review*. **69**: 520-536.
- Ronning, R.R., Glover, J.A., Conoley, J.C. & Witt, J.C. (1987). *The influence of cognitive psychology on testing*. Hillsdale, NJ: Erlbaum.
- Royer, F.L. (1971). Information processing of visual figures in the Digit Symbol Substitution Task. *Journal of Experimental Psychology*, **87**: 344-342.
- Shepard, R.N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, **171**: 701-703.
- Sternberg, R.J. (1977). *Intelligence, information processing, and analogical reasoning: the componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.
- Sternberg, R.J. (Ed.). (1985). *Human abilities: an information-processing approach*. New York: Freeman.
- Sternberg, S. (1966). High speed scanning in human memory. *Science*, **153**: 652-654.
- Tapsfield, P.G.C. (1993a). *The British Army Recruit Battery. Test-Retest Reliability*. HAL Technical Report: 5-1993 (APRE).
- Tapsfield, P.G.C. (1993b). *The British Army Recruit Battery. 1993 Applicant Norms*. HAL Technical Report: 6-1993 (APRE).
- 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).
- Tatsuoka, K. M., & Tatsuoka, M. M. (1978). Time-score analysis in criterion-referenced tests. Report of the Computer-Based Education Research Laboratory (CERL Report E-1), University of Illinois, Urbana, Ill.
- Thurstone, L.L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press and latterly as Medway Reprint (1975) University of Chicago Press.
- Vernon, P.A. (1983). Speed of information processing and general intelligence. *Intelligence*, **7**: 53-70.
- Vernon, P.A. & Jensen, A.R. (1985). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences*, **5**: 411-423.
- Wainer, H. & Messick, S. J. (Eds.), (1983). *Principals of modern psychological measurement*. Hillsdale, NJ: Erlbaum.

- Wainer H. (2002), On the automatic generation of test items: Some whens, whys and hows. In Irvine, S.H. & Kyllonen, P.C. (Eds.). *Item generation for test development*. Erlbaum Associates, Mahwah, NJ pp.287-316.
- Weibull, W. (1951). A statistical distribution of wide application. *Journal of Applied Mechanics*, 1951, **18**: 293-297.
- White, P.O. (1982). Some major components in general intelligence. In H.J. Eysenck, (Ed.), *A model for intelligence*. Springer-Verlag: New York.
- Woltz, D.J. (1987). *Activation and decay of semantic memory: An individual differences investigation of working memory*: MS submitted for publication. Brooks AFB, Texas: Manpower & Personnel Division, Air Force Human Resources Laboratory.
- Wright, D.E. (1990). *Item Response and Theory for Item Generation: Comment and Developments*. HAL Technical Report: 3-1990 (APRE).
- Wright, D.E. (1992). *IRT Modelling Using Latent Variable Generalised Linear Models*. HAL Technical Report: 3-1992 (APRE).
- Wright, D.E. & Dennis, I. (1992). *Development of a test of mental cube folding for use in officer selection*. Technical Report for Science Air (3), Ministry of defence, London. Human Assessment Laboratory, University of Plymouth, Plymouth, UK.
- Wright, D.E, Irvine, S.H. & Tapsfield, P.G.C. (1992). *Test Lengths and Reliabilities*. HAL Technical Report: HAL1-1992 (APRE).
- Wright, D. E. (2002). Scoring tests when items have been generated. In Irvine, S.H. & Kyllonen, P.C. (Eds.). *Item generation for test development*. Erlbaum Associates, Mahwah, NJ pp.277-286.

Annexe A: Description of Measures Used

The Tests of Basic Skills

ALPHABET TEST (AB)¹

The *Alphabet Test* presents a short list of common German surnames in each item. They have to be put in the correct alphabetic order. This is done by clicking on them in turn. Errors are indicated and have to be corrected. Scores are given for fluency and accuracy. Employers are right in thinking that fluent use of the alphabet is an essential part of literacy.

ERROR DETECTION (ED)

The *Error Detection Test* measures the speed with which features of letters and numerals are perceived to be the same or different, and the detection of mismatches. This quality - Perceptual Speed - is a recognised building block in the development of literacy skills and clerical accuracy.

NUMBER FLUENCY (NF)

The *Number Fluency test* is designed to simulate work tasks that require elementary number skills unaided by calculators, or close attention to number precision in computing situations. Consequently, the test has a very high attentional demand. Attention lapses result in slow work or errors. People who are easily distracted will not do as well as others who can attend closely. Test items require only basic numeracy in addition, subtraction, multiplication and division.

The Working Memory Tests

Working memory is concerned with the immediate processing of new information with sufficient reliability to allow a new action to follow. Working memory tests predict individual differences in a wide spectrum of jobs and training contexts where sustained mental effort is the key to performance. Working memory tasks have proved to be among the most consistent predictors of learning new skills.

ODDS AND EVENS (OE)

Odds and Evens is a test of comparisons. It is easy to understand. The test requires the correct identification of odd and even number sequences, given a set of rules for the order of three numbers. One must compare the order of rules with the actual order of numbers and decide how many match.

The test is designed to represent work tasks that require normal working memory skills unaided by look-up tables or instructional manuals. It is an important component of general intelligence.

WORD RULES (WR)

The *Word Rules* test is a verbal version of the *Odds and Evens* test just described: and the formats are identical. It is used whenever more emphasis on verbal comprehension is required by the job description. The test requires the correct identification of word sequences given a set of rules for their order. Because it has vocabulary common to the *Reasoning Categories* test, these two tests are not given together.

The Verbal Aptitude Tests

The verbal tests are powerful, but they are constructed from simple sentences so that the threshold of understanding is not a barrier to testing the essentials. These tests concentrate on following instructions, learning rules for processing information and drawing conclusions.

REASONING CATEGORIES (RC)

Good performance on the *Reasoning Categories* test depends on memorising a list of categories, such as *building* and *tree*, and putting it in a prescribed order. Then the order has to be matched exactly with specific examples, such as *school* and *pine* to answer the question. It is designed to simulate work tasks that require memory for a number of 'if-then' instructions or procedures. People who can extract essentials from written instructions (for example software, technical and 'rule-book' manuals) and act upon them will score highly on this test. This is the longest test in the series. It tends to be reserved for higher-level job applicants.

DEDUCTIVE REASONING (IT)

This is a short low-literacy level of test *Deductive Reasoning*, in the well-known *Transitive Inference* format. The *Transitive Inference* tasks have been proved wherever work functions demand simple deductions from a small amount of information. The literacy threshold is low, but the mental demands are significant in each of two variants, standard and advanced.

The Visualisation Tests

ORIENTATION (OR)

Spatial orientation tests, particularly those involving the mental rotation of shapes, have been used for over 70 years. They screen applicants for jobs where visualisation is a requirement, for example in building, engineering, design, architecture photography and other technical roles. Persons with high aptitude for technical training out-perform those for whom the path

¹ The abbreviations used in the system are all listed in brackets for reference.

to technical excellence has proved longer and more difficult. The test has particular promise for identifying potential navigators and air-traffic controllers.

The *Orientation Test* is based on the relative position of two arrows of different colours that point in one of four directions. The orientation of two arrows is verbally described. This description has to be transformed mentally to a figural representation, and then visually identified in one of eight possible solutions. It is used in technical and executive testing programs. In the computer delivered version, the answers are shown on a separate screen to increase the power of the test.

Computer Delivery

Computer delivery, scoring and report mechanisms are automatic. The standard computer delivered test interface is by mouse with a restricted cursor. Mouse practice is provided for each participant. All tests have graduated instructions and a final set of five practice items. If fewer than four are answered correctly, the practice items are repeated again. Progression from the second practice set to the test itself is automatic.

The Bundeswehr Entry Tests and Other Measures

A number of measures were provided by the German Ministry of Defence, including test centre identification number, results of aptitude tests, and age. No participant was identified by name or country of origin.

The Bundeswehr Entry Tests are administered by computer using a console interface. Some aptitude tests are adaptive. Other tests are timed, having fixed numbers of trials at a specific task.

Tests are scored either by estimating ability thresholds through adaptive formulas, or by raw scores that are then transformed by a uniform metric. Theta is the ability parameter estimated with three adaptive tests of Verbal Reasoning, Mathematical Facility and Figural Reasoning. A general aptitude parameter (or intelligence), *int_theta* is the sum of those three thetas. All four thetas have been standardised according to the formula $4 - 4/3 z$; hence *int_note*, *vr_note*, *num_note* and *fr_note*. This is the same formula used to transform all scores. The subscript *_note* always denotes a standardised score.

int_theta int_note

Intelligence Test Theta and Standard Score/Value (Note). This composite consists of adaptively (2-PL Birnbaum) administered items out of three different item pools (figural reasoning, number and verbal reasoning/comprehension). This scale is stretched, using 0,2 steps from 1,0..1,2...to 6,8...7,0 (also scale mean 4,0 and SD $-4/3$).

vr_theta vr_note fr_theta fr_note

Verbal Reasoning and Figural Reasoning Scores. See above, but integers used for standardised values 1 to 7.

num_theta num_note

Mathematical Facility Scores. The item pool consists of geometrical, arithmetic problems, numerical comprehension tasks. It is not a number fluency test, however, and it is not speeded.

mtr_scor mtr_note

Mechanical Comprehension Scores. These are derived from a number of items demanding understanding of mechanical principles.

rpr_scor rpr_note

Reaction Time Test. The participant has to press 0, 1 or 2 out of 6 keys on the console (Annexe C), depending on which (or none) of 6 stimuli have been presented. The response is correct if all stimuli have been recognised. Those stimuli are a triangle, a square, a circle, a cross, a high pitch and a low pitch sound.

rst_scor rst_note

Orthographic Test. This is a multiple-choice test of correct word identification from four different spellings of the same word.

ekt_scor ekt_note

Electronic Comprehension Test. Principles and applications of electrical engineering, e.g. Ohm's law.

fr_scor fr_note

Radio Operator Test. The participant has to discriminate the letters R, W, K presented acoustically as Morse signals. After being trained, the participant has to run three trials of 50 signals each, the signals being presented faster from trial to trial. This is a highly speeded test; a maximum of 150 Items administered by headphones.

dop_scor dop_note

Auditory Discrimination Test. Two sounds are presented one after the other. The participant has to decide which one was higher; or if both have the same pitch. They are mono sounds presented on both ears. The differences between those two sounds range from 32 Hz to 5 Hz, the pitches range from 820 Hz to 940 Hz.

sig_scor sig_note

Signal Detection Test. Four optical stimuli (flashing circles, short flash, long flash) are presented. The participant has to reproduce the order of those short or long stimuli: (s l s l; or l l l s for example) by pressing keys on the panel. There are 20 short/long patterns of light signals to detect and encode this pattern by pressing buttons on the panel. The short signals last 200 ms; the long signals 700 ms; the time between the signals is 300 ms. (highly speeded test).

Other Categories and Measures Used in the Analysis

Because of the limited number of subjects available, every care had to be exercised in the partition of variance between parallel test forms. The following categories and variables were used in the relevant analyses.

Test Forms

Three parallel forms of each of *The Bundeswehr Experimental Test Battery* tests were administered at random to each participant. A *Test Form* variable was created to identify each form of every test administered. *Test Form* was used as a random factor in all ANOVAs.

Filter Variable of Test Competence

Of particular note is a filter variable used to exclude from calculations of anchor norms and reliability estimates those subjects who had a single chance score in any of the item-generative tests. This category excludes participants who scored less than zero on any one of the eight experimental tests – a sign of failure to comprehend instructions in German, or failure to carry out the instructions properly. The filter variable is called *Test Competence*. These measures and categories ensured a database of subjects for whom one might assume that the tests were understood and completed satisfactorily.

Test Competence was also used as a category in assessing test parallelism variation. Two groups were used: those who had no score below zero or a minimum threshold (all positive); and those who had one or more negative scores (one or more negative).

Age

Although conscripts comprise a population with a very restricted age range, in previous analyses for the progress report (Irvine et al, 2003) age was found to correlate negatively with test performance. Younger participants were better performers than older ones. Consequently, age is a necessary covariate in general linear analysis of variance models for assessing test form effects.

Educational Level

Level of education was recorded for participants. Three finite groups were identified: Hauptschulabschluss (Group1); Realschulabschluss (Group 2); and (Allgemeine Hochschulreife Group3). Also included in Group 3 were participants with University/College entry qualifications.