GENERAL, CRITICAL & GRADUATE TEST BATTERY

the technical manual

8

Test Batteries

PSYTECH

CONTENTS

②

LIST OF TABLES

④

1 THEORETICAL OVERVIEW

A major reason for using psychometric tests to aid selection decisions is that they provide information that cannot be obtained easily in other ways. If such tests are not used then what we know about the applicant is limited to the information that can be gleaned from an application form or CV, an interview and references. If we wish to gain information about a person's specific aptitudes & abilities and about their personality, attitudes and values then we have little option but to use psychometric tests. In fact, psychometric tests can do more than simply provide additional information about the applicant. They can add a degree of reliability and validity to the selection procedure that it is impossible to achieve in any other way. How they do this is best addressed by examining the limitations of the information obtained through interviews, application forms and references and exploring how some of these limitations can be overcome by using psychometric tests.

⑥

THE ROLE OF PSYCHOMETRIC TESTS IN PERSONNEL SELECTION AND ASSESSMENT

While much useful information can be gained from the interview, which clearly has an important role in any selection procedure, it does nonetheless suffer from a variety of weaknesses. Perhaps the most important of these is that the interview as been shown to be a very unreliable way to judge a person's character. This is because it is an unstandardised assessment procedure. That is to say, each interview will be different from the last. This is true even if the interviewer is attempting to ask the same questions and act in the same way with each applicant. It is precisely this aspect of the interview that is both its main strength and its main weakness. The interview enables us to probe each applicant in depth and discover individual strengths and weaknesses. Unfortunately, the interviews unstandardised, idiosyncratic nature makes it difficult to compare applicants, as it provides no base line against which to contrast interviewees' differing performances. In addition, it is likely that different interviewers may come to radically different conclusions about the same applicant. Applicants will respond differently to different interviewers, quite often saying very different things to them. In addition, what any one applicant might say will be interpreted quite differently by each interviewer. In such cases we have to ask which interviewer has formed the correct impression of the candidate? This is a question to which there is no simple answer.

A further limitation of the interview is that it only assesses the candidate's behaviour in one setting, and with regard to a small number of people. How the candidate might act in different situations and with different people (e.g. when dealing with people on the shop floor) is not assessed, and cannot be predicted from an applicant's interview performance. Moreover, the interview provides no reliable information about the candidate's aptitudes and abilities. The most we can do is ask the candidate about his strengths and weaknesses, a procedure that has obvious limitations. Thus the range, and reliability of the information that can be gained through an interview are limited.

There are similar limitations on the range and usefulness of the information that can be gained from application forms or CV's. While work experience and qualifications may be prerequisites for certain occupations, in and of themselves they do not determine whether a person is likely to perform well or badly. Experience and academic achievement is not always a good predictor of ability or future success. While such information is important it may not be sufficient on its own to enable us to confidently choose between applicants. Thus aptitude and ability tests are likely to play a significant role in the selection process as they provide information on a person's potential and not just their achievements to date.

Moreover, application forms tell us little about a person's character. It is often a candidate's personality that will make the difference between an average and an outstanding performance. This is particularly true when candidates have relatively similar records of achievement and past performance. Therefore, personality tests can play a major role in assisting selection decisions.

References do provide some useful information but mainly for verification purposes. While past performance is undoubtedly a good predictor of future performance references are often not good predictors of past performance. If the name of the referee is supplied by the applicant, then it is likely that they have chosen someone they expect to speak highly of them. They will probably have avoided supplying the names of those who may have a less positive view of their abilities. Aptitude and ability tests, on the other hand, give us an indication of the applicant's probable performance under exam conditions. This is likely to be a true reflection of the person's ability.

What advantages do psychometric tests have over other forms of assessment? The first advantage they have is that they add a degree of reliability to the selection procedure that cannot be achieved without their use. Test results can be represented numerically making it easy both to compare applicants with each other, and with pre-defined groups (e.g. successful vs. unsuccessful job incumbents). In the case of personality tests the test addresses the issue of how the person characteristically behaves in a wide range of different situations and with different people. Thus psychometric tests of personality, aptitude and ability provide a range of information that are not easily and reliably assessed in other ways. Such information can fill important gaps which have not been assessed by application forms, interviews and references. It can also raise questions that can later be directly addressed in the interview. It is for this reason that psychometric tests are being used increasingly in personnel selection. Their use adds a degree of breadth to assessment decisions which cannot be achieved in any other way.

8

THE ORIGINS OF REASONING TESTS

The assessment of intelligence or reasoning ability is perhaps one of the oldest areas of research interest in psychology. Gould (1981) has traced attempts to scientifically measure psychological aptitudes and abilities to the work of Galton at the end of the last century. Prior to Galton's pioneering work, however, interest in this area was aroused by phrenologists' attempts to assess mental ability by measuring the size of people's heads. Reasoning tests, in their present form, were first developed by Binet, a French educationalist who published the first test of mental ability in 1905.

Binet was concerned with assessing the intellectual development of children and to this end invented the concept of mental age. Questions, assessing academic ability, were graded in order of difficulty according to the average age at which children could successfully complete each item. From the child's performance on such a test it was possible to derive its mental age. This involved comparing the performance of the child with the performance of the 'average child' from different age groups. If the child performed at the level of the average 10 year old, then the child was said to have a mental age of 10, regardless of its chronological age. From this idea the concept of the Intelligence Quotient (IQ) was developed by William Stern (1912) who defined it as mental age divided by chronological age multiplied by 100. Previous to Stern's paper chronological age had been subtracted from mental age to provide a measure of mental alertness. Stern showed that it was more appropriate to take the ratio of these two constructs, which would provide a measure of the child's intellectual development relative to other children. He further proposed that this ratio should be multiplied by 100 for ease of interpretation; thus avoiding cumbersome decimals.

Binet's early tests were subsequently revised by Terman et al. (1917) to produce the now famous Stanford-Binet IQ test. These early IQ tests were first used for selection by the American's during the first world war, when Yerkes (1921) tested 1.75 million soldiers with the army alpha and beta tests. Thus by the end of the war, the assessment of reasoning ability had firmly established its place within psychology.

2 THE GRADUATE & GENERAL REASONING TESTS (GRT1 & GRT2)

Research in the area of intelligence testing has consistently demonstrated that three aptitude domains: Verbal, Numerical and Abstract Reasoning Ability (Heim, 1970). Consequently, the GRT1 and GRT2 have been designed to measure just these three areas of ability. Verbal and Numerical ability assess, as their respective names would suggest, the ability to use words and numbers in a rational way, correctly identifying logical relationships between these entities and drawing conclusions and inferences from them. Abstract, reasoning assesses the ability to identify logical relationships between abstract spatial relationships and geometric patterns. Many psychologists argue that Abstract Reasoning tests assess the purest form of 'intelligence'. That is to say, these tests are the least affected by educational experience and assess what some might term 'innate' reasoning ability. Namely, the ability to solve abstract, logical problems which require no prior knowledge or educational experience.

Research has clearly demonstrated that in order to accurately assess reasoning ability it is necessary to develop tests which have been specifically designed to measure that ability in the population under consideration. That is to say, we need to be sure that the test has been developed for use on the particular group being tested, and thus that it is appropriate for that particular group. There are two ways in which this is important. Firstly, it is important that the test has been developed in the country in which it is intended to be used. This ensures that the items in the test are drawn from a common, shared cultural experience, giving each candidate an equal opportunity to understand the logic which underlies each item. Secondly, it is important that the test is designed for the particular ability range on which it is to be used. A test designed for those of average ability will not accurately distinguish between people of high ability as all their scores will cluster towards the top end of the scale. Similarly, a test designed for people of high ability will be of little practical use if given to people of average ability. Not only will the test not discriminate between applicants, as all the scores will cluster towards the bottom of the scale, but also as the questions will be too difficult for most of the applicants they are likely to lose motivation, producing artificially low scores. For this reason two versions of this reasoning test were developed. One developed for the general population (the average ability range) and one for the graduate population.

In constructing the items in the GRT1 and GRT2 which measure these reasoning abilities a number of guide lines were borne in mind. Firstly, and perhaps most importantly, each item was constructed so that only a minimal educational level was needed in order to be able to correctly solve each item. Thus we tried to ensure that each item was a measure of 'reasoning ability', rather than being a measure of specific knowledge or experience. For example, in the case of the numerical items the calculations involved in solving each item are relatively simple, with the difficulty of the item being due to the logic which underlies that question rather than being due to the need to use complex mathematical operations to solve that item. (It should be noted however that in the case of the GRT1 a higher level of education was assumed, as the test was designed for a graduate population). Secondly, a number of different item types (e.g. odd one out, word meanings etc.) were used to measure each aspect of reasoning ability. This was done in order to ensure that each sub-scale measures a broad aspect of reasoning ability (e.g. Verbal Reasoning Ability), rather than measuring a very specific aptitude (e.g. vocabulary). In addition, the use of different item types ensures that the test is measuring different components of reasoning ability. For example the ability to understand analogies, inclusion/exclusion criteria for class membership etc.

⑫

3 THE CRITICAL REASONING TEST BATTERY (CRTB)

Research has clearly demonstrated that in order to accurately assess reasoning ability it is necessary to develop tests which have been specifically designed to measure that ability in the population under consideration. That is to say, we need to be sure that the test has been developed for use on the particular group being tested, and thus is appropriate for that particular group. There are two ways in which this is important. Firstly, it is important that the test has been developed in the country in which it is intended to be used. This ensures that the items in the test are drawn from a common, shared cultural experience, giving each candidate an equal opportunity to understand the logic which underlies each item. Secondly, it is important that the test is designed for the particular ability range on which it is to be used. A test designed for those of average ability will not accurately distinguish between people of high ability as all the scores will cluster towards the top end of the scale. Similarly, a test designed for people of high ability will be of little use if given to people of average ability. Not only will it not discriminate between applicants, as all the scores will cluster towards the bottom of the scale, but also as the questions will be too difficult for most of the applicants they are likely to be de-motivated, producing artificially low scores. Consequently, the VCR1 and NCR1 have been developed on data from undergraduates. That is to say, people of above average intelligence, who are likely to find themselves in senior management positions as their career develops.

In constructing the items in the VCR1 and NCR1 a number of guide lines were borne in mind. Firstly, and perhaps most importantly, special care was taken when writing the items to ensure that in order to correctly solve each item it was necessary to draw logical conclusions and inferences from the stem passage/table. This was done to ensure that the test was assessing critical (logical/deductive) reasoning rather than simple verbal/numerical checking ability. That is to say, the items assess a person's ability to think in a rational, critical way and make logical inferences from verbal and numerical information, rather than simply check for factual errors and inconsistencies.

In order to achieve this goal for the Verbal Critical Reasoning (VCR1) test two further points were born in mind when constructing the stem passages for the VCR1. Firstly, the passages were kept fairly short and cumbersome grammatical constructions were avoided, so that a person's scores on the test would not be too affected by reading speed; thus providing a purer measure of critical reasoning ability. Secondly, care was taken to make sure that the passages did not contain any information which was counter-intuitive, and was thus likely to create confusion. To increase the acceptability of the test to applicants the themes of the stem passages were chosen to be relevant to a wide range of business situations. As a consequence of these constraints the final stem passages were similar in many ways to the short articles found in the financial pages of a daily newspaper.

Finally an extended response format was chosen for the VCR1. While many critical reasoning tests only ask the applicant to make the distinction between whether the target statement is true, false or cannot be inferred from the stem passage the response format for the VCR1 was extended to include the judgement of whether the target statement was probably or definitely true/false given the information in the passage. This was done in order to decrease the chance of guessing a correct answer from 33% to 25%. With guessing having substantially less impact on a candidate's final score, it was thus possible to decrease the number of items in the test that were needed for it to be reliable.

⑯

4 THE PSYCHOMETRIC PROPERTIES OF THE REASONING TESTS

This chapter will present details concerning the psychometric properties of the reasoning Tests. The aim will be to show that these measures fulfil various technical requirements, in the areas of standardisation, reliability and validity, which ensure the psychometric soundness of the test.

Standardisation : Normative

Normative data allows us to compare an individuals score on a standardised scale against the typical score obtained from a clearly identifiable, homogeneous group of people.

RELIABILITY

The property of a measurement which assesses the extent to which variation in measurement is due to true differences between people on the trait being measure or to measurement error.

In order to provide meaningful interpretations, the reasoning tests were standardised against a number of relevant groups. The constituent samples are fully described in the next section. Standardisation ensures that the measurements obtained from a test can be meaningfully interpreted in the context of a relevant distribution of scores. Another important technical requirement for a psychometrically sound test is that the measurements obtained from that test should be reliable.

Reliability is generally assessed using two specific measures, one related to the stability of scale scores over time, the other concerned with the internal consistency, or homogeneity of the constituent items that form a scale score.

Reliability : Stability

Also known as test-retest reliability, an assessment is made of the similarity of scores on a particular scale over two or more test occasions. The occasions may be from a few hours, days, months or years apart. Normally Pearson correlation coefficients are used to quantify the similarity between the scale scores over the two or more occasions.

Stability coefficients provide an important indicator of a test's likely usefulness of measurement. If these coefficients are low (< approx. 0.6) then it is suggestive of either that the abilities/behaviours/attitudes being measured are volatile or situationally specific, or that over the duration of the retest interval, situational events have made the content of the scale irrelevant or obsolete. Of course, the duration of the retest interval provides some clue as to which effect may be causing the unreliability of measurement. However, the second measure of a scales reliability also provides valuable information as to why a scale may have a low stability coefficient. The most common measure of internal consistency is Cronbach's Alpha. If the items on a scale have high inter-correlations with each other, and with the total scale score, then coefficient alpha will be high. Thus a high coefficient alpha indicates that the items on the scale are measuring very much the same thing, while a low alpha would be suggestive of either scale items measuring different attributes or the presence of error.

Reliability : Internal Consistency

Also known as scale homogeneity, an assessment is made of the ability of the items in a scale to measure the same construct or trait. That is a parameter can be computed that indexes how well the items in a scale contribute to the overall measurement denoted by the scale score. A scale is said to be internally consistent if all the constituent item responses are shown to be positively associated with their scale score.

The fact that a test has high internal consistency & stability coefficients only guarantees that it is measuring something consistently. It provides no guarantee that the test is actually measuring what it purports to measure, nor that the test will prove useful in a particular situation. Questions concerning what a test actually measures and its relevance in a particular situation are dealt with by looking at the tests validity. Reliability is generally investigated before validity as the reliability of test places an upper limit on tests validity. It can be mathematically demonstrated that a validity coefficient for a particular test can not exceed that tests reliability coefficient.

VALIDITY

The ability of a scale score to reflect what that scale is intended to measure. Kline's (1993) definition is "A test is said to be valid if it measures what it claims to measure".

Validation studies of a test investigate the soundness and relevance of a proposed interpretation of that test. Two key areas of validation are known as criterion validity and construct validity.

Validity : Criterion Validity

Criterion validity involves translating a score on a particular test into a prediction concerning what could be expected if another variable was observed.

The criterion validity of a test is provided by demonstrating that scores on the test relate in some meaningful way with an external criterion. Criterion validity comes in two forms – predictive & concurrent. Predictive validity assesses whether a test is capable of predicting an agreed criterion which will be available at some future time – e.g. can a test predict the likelihood of someone successfully completing a training course. Concurrent validity assesses whether the scores on a test can be used to predict a criterion measure which is available at the time of the test – e.g. can a test predict current job performance.

Validity : Construct Validity

Construct validity assesses whether the characteristic which a test is actually measuring is psychologically meaningful and consistent with the tests definition.

The construct validity of a test is assessed by demonstrating that the scores from the test are consistent with those from other major tests which measure similar constructs and are dissimilar to scores on tests which measure different constructs.

STANDARDISATION

For each of the three reasoning batteries, information is provided on the constituent norm samples, age and gender differences where they apply. All normative data is available from within the GeneSys system which computes for any given raw score, the appropriate standardised scores for the selected reference group. In addition the GeneSys™ software allows users to establish their own in-house norms to allow more focused comparison with profiles of specific groups.

GRT2 NORMATIVE DATA

The total norm base of the GRT2 is based on a general population norm as well as a number of more specialised norm groups. These include undergraduates, technical staff, personnel managers, customer service staff, management applicants etc. detailed in Table 1.

GRT2 GENDER AND AGE DIFFERENCES

Gender differences on GRT2 were examined by comparing results of almost equal numbers of males and female respondents matched as far as possible for educational and socio-economic status. Table 2 provides mean scores for males and females on each of the GRT2 scales as well as the t-value for mean score differences.

The results below demonstrate gender differences on NR2 with the male mean score just over two raw score points higher than that of the females. This is in line with other numerical measures. More surprising perhaps, is that no differences were observed on either the Verbal and Abstract measures.

The effect of age on GRT2 scores was examined using a sample of 1441 respondents on whom age data was available (see Table 3). Whereas there is a negative relationship with GRT2 scores and age for all the three measures, this tendency is highly significant on the Abstract (AR2), suggesting, in line with expectations that fluid ability may be more likely to decline with age than crystallised ability.

	Males N	Females N	Mean Age	Range	SD Age
General Population	3177	1236	35.10	16-63	8.87
Telesales Applicants	175	391	27.16	16-55	8.29
HE College Students	5	158	16.25	15-41	3.14
Customer Service Clerks	28	87	26.46	19-50	7.19
Technical Staff	122	24	29.38	16-58	9.92
Financial Consultants	58	20	32.94	20-53	9.20
HR Professionals	30	38	36.31	22-55	7.94
Service Engineers	86	8	26.56	18-58	6.87

Table 1: GRT2 Standardisation Samples

	Mean	Mean				Females	Males
GRT2	**Females**	**Males**	**t-value**	**df**	**p**	**N**	**N**
VR2	23.36	23.22	.3404	1438	.734	852	588
NR2	15.77	17.91	-5.424	1438	.000	852	588
AR2	16.95	17.40	-1.259	1438	.208	852	588

Table 2: Gender differences on GRT2

GRT2	AGE
VR2	.11
NR2	.10
AR2	-.37

Table 3: Pearson correlations between GRT2 and age

GRT1 GENDER AND AGE DIFFERENCES

Gender differences on GRT1 were examined by comparing results of almost equal numbers of males and female respondents matched as far as possible for educational and socio-economic status. Table 4 provides mean scores for males and females on each of the sub-scales of the GRT1 as well as the t-value for mean score differences.

Consistent with findings on the GRT2, the only significant score difference is registered on the Numerical (NR1) with males registering a higher mean score. No significant differences were observed for either the Verbal or the Abstract components of the GRT1.

The effect of age on GRT1 scores was examined using a sample of 499 respondents on whom age data was available (see Table 5). Whereas the Verbal and Numerical were found to be unrelated to age, the Abstract showed a significant negative relationship, consistent with expectations. The correlations are lower than those obtained with the GRT2, although this may be explained by the sample which in this case is more restricted in the range of observed scores.

CRTB GENDER AND AGE DIFFERENCES

Gender differences on CRTB were examined by comparing results of males and female respondents matched as far as possible for educational and socio-economic status. Table 6 opposite provides mean scores for males and females on both the verbal and numerical sub-scale of the CRTB scales as well as the t-value for mean score differences.

While female respondents register marginally higher verbal reasoning scores (VCR1) than males, this is not statistically significant. A significant gender difference was observed on the Numerical (NCR1), with males registering over four raw score points higher mean score. This is consistent with other measures of numerical ability.

The effect of age on CRTB scores was examined using a sample of 359 respondents on whom age data was available (see Tabel 7). Unlike the General and Graduate Reasoning tests, the correlations with age are positive for the Critical Reasoning, although only marginal and non-significant for the Verbal. This suggests that the Numerical at least, may be measuring more of an acquired ability, which is more positively influenced by experience than other more classic measures of numerical ability.

GRT1	Mean Females	Mean Males	t-value	df	p	Females N	Males N
VR1	16.73	16.61	.260	497	.795	251	248
NR1	12.98	14.91	-4.255	497	.000	251	248
AR1	14.11	14.27	-.383	497	.702	251	248

Table 4: Gender Differences on GRT1

GRT1	AGE
VR1	-.00
NR1	-.05
AR1	-.27

Table 5: Relationship between Age and GRT1

CRTB	F	M	t-value	df	p	F	M
VCR1	18.34	17.78	.501	134	.6178	47	89
NCR1	10.19	14.40	-3.578	134	.0004	47	89

Table 6: Gender differences on CRTB

CRTB	AGE
VRC1	.03
NRC1	.18

Table 7: Relationship between Age and CRTB

RELIABILITY OF THE REASONING TESTS

If a reasoning test is to be used for selection and assessment purposes the test needs to measure each of the aptitude or ability dimensions it is attempting to measure reliably, for the given population (e.g. graduate entrants, senior managers etc.). That is to say, the test needs to be consistently measuring each ability so that if the test were to be used repeatedly on the same candidate it would produce similar results.

It is generally recognised that reasoning tests are more reliable than personality tests and for this reason high standards of reliability are usually expected from such tests. While many personality tests are considered to have acceptable levels of reliability if they have reliability coefficients in excess of .7, reasoning tests should have reliability coefficients in excess of .8.

GRT2 INTERNAL CONSISTENCY

Table 8 presents alpha coefficients for the three sub-scales of the GRT2 (n=135). Each of these reliability coefficients is substantially greater than .8, clearly demonstrating that general population version of the GRT is highly reliable.

GRT1 INTERNAL CONSISTENCY

Table 9 presents alpha coefficients for the three sub-scales of the GRT1 (n=109). Each of these reliability coefficients is greater than .8, clearly demonstrating that the graduate version of the GRT meets acceptable levels of internal consistency.

GRT1 TEST-RETEST

A sample of 70 undergraduate students completed the GRT1 on two separate occasions with a four week interval. Table 10 provides uncorrected correlations for each measure.

Although the Abstract falls somewhat below the ideal, the test-retest correlations are generally of a high and acceptable level which, in conjunction with the internal consistency data would demonstrate that the GRT1 is a reliable measure of general reasoning ability.

Table 11 presents alpha coefficients for the two sub-scales of the Critical Reasoning Test. Each of these reliability coefficients is .8 or greater, clearly demonstrating that Critical Reasoning Tests reach acceptable levels of reliability.

GRT2	r
Verbal (VR2)	.83
Numerical (NR2)	.84
Abstract (AR2)	.83

Table 8: Coefficient Alpha for GRT2 Sub-scales (n = 135)

GRT1	r
Verbal (VR1)	.82
Numerical (NR1)	.85
Abstract (AR1)	.84

Table 9: Coefficient Alpha for GRT1 Sub-scales (n = 109)

GRT1	VR1	NR1	AR1
VR1	**.79**	.31	.09
NR1	.35	**.78**	.39
AR1	.22	.38	**.74**

Table 10: Test-retest reliability estimates for GRT1 (N=70)

CRTB	r
Verbal Critical Reasoning (VCR1)	.80
Numerical Critical Reasoning (NCR1)	.82

Table 11: Coefficient Alpha for CRTB Sub-scales (N=134)

METHOD EFFECTS

One important aspect of the change from paper and pencil to computer-based administration, concerns the effects the change in test format might have on the available normative data for a test. In other words does the translation of a test to computer format alter the nature of the test itself? It is possible that the score range of a test administered in question booklet form may be different from the range when the test is administered via a computer. Given that most of the normative data available for mainstream psychometric tests was collected from paer and pencil test administration the question is far from academic. For instance, many organisations will administer computer-based tests at their head office location but will use paer and pencil format when their testers visit remote locations. If comparison is to be made across a group in which some individuals received paer and pencil testing and some computer testing then the importance of the question of normative comparability can not be overstated.

Very little research attention has been paid to this topic, which is surprising given the possible impact that format differences would have. Roid (1986) has indicated that what little evidence is available concerning paer and pencil v. computer formats suggests that computer administration of most tests does not change the score range enough to affect the normative basis of the test. While this is somewhat reassuring the number of studies looked at by Roid was small and consisted entirely of American investigations.

As publishers of a wide range of tests which can be administered both by paer and pencil and computer, Psytech International recently conducted a study looking at the effect of administration format on reasoning test performance. The Test chosen was the Graduate Reasoning Test, a graduate level test comprising three sub-tests – verbal, numerical and abstract. This test was chosen as being representative of many mainstream reasoning tests. The Graduate Reasoning Test also gives an opportunity to test whether the representational format of a test is important. It could be the case that computer presentation of graphical material, such as is found in mechanical and spatial reasoning tests, might lead to performance differences while alphanumeric material does not. With the GRT1 the abstract subtest uses a graphical format while the verbal and numerical sub-scales use alphanumeric formats.

THE STUDY

A group of 80 university undergraduates took part in this investigation. The students were divided into four groups. Each group completed the Graduate Reasoning Test twice with an interval of two weeks between successive administrations. Two of the groups completed the same format of the test on each occasion – i.e. paper/paper or computer/computer. The other two groups experienced both formats of the test, each group in a different order – i.e. paper/computer or computer/paper.

RESULTS

An independent t-test was used to test whether any differences existed between the mean scores on first administration for those students who completed the paper version of the GRT1 and those receiving the GeneSys Administered version. As can be seen from Table 12 a significant difference was found for both the Numerical and Abstract subscales in that students who had received the computer administered version of the test tended to score significantly higher than those who had received the paper version first. Table 12 provides similar data for the two test formats on second administration. It can be seen from this table that no significant differences existed between the different formats for second administration.

CONCLUSION

The results of this study show that as far as the Graduate Reasoning Test, at least, is concerned the format in which the test is administered does affect, to some extent, the scores obtained. No differences between the group means were detected for either of the three sub-tests of the Graduate Reasoning Test. It was also the case that no difference was found between the graphical representation of the abstract sub-test and the alpha-numerical representation of the verbal and numerical sub-tests. Furthermore no interaction effects were observed which indicates that the well established phenomenon of 'practice effects' does not differ with the nature of the test medium.

These results provide some confidence that the changeover from paper and pencil to computer-based testing will not require the re-standardisation of tests. It would seem from this study that administration of ability tests by either computer or paer and pencil will produce similar performance levels. Thus it is perfectly acceptable to compare individuals who were tested using different formats.

GRT1	Mean PC	Mean Paper	t-value	df	p	N-PC	N-Paper	SD-PC	SD-Paper	F-ratio variance	P variance
VR1	19.09	18.48	.876	158	.382	80	80	4.26	4.58	1.15	.532
NR1	14.53	13.86	.927	158	.355	80	80	3.97	5.01	1.59	.0419
AR1	15.58	14.79	1.117	158	.266	80	80	3.81	5.03	1.74	.0146

Table 12: Differences between Computer vs. Conventional Methods of Test Administration

VALIDITY

Whereas reliability assess the degree of measurement error of a reasoning test, that is to say the extent to which the test is consistently measuring one underling ability or aptitude, validity addresses the question of whether or not the scale is measuring the characteristic it was developed to measure. This is clearly of key importance when using a reasoning test for assessment and selection purposes. In order for the test to be a useful aid to selection we need to know that the results are reliable and that the test is measuring the aptitude it is supposed to be measuring. Thus after we have examined a test's reliability we need to address the issue of validity. We traditionally examine the reliability of a test before we explore its validity as reliability sets the lower bound of a scale's validity. That is to say a test cannot be more valid than it is reliable.

GRT1 sub-scale	VR1	NR1	AR1
Verbal (VR1)	_	.42	.30
Numerical (NR1)		_	.55
Abstract (AR1)			_

Table 13: Product-moment Correlations between the GRT1 Sub-scales (n=499)

GRT2 sub-scale	VR2	NR2	AR2
Verbal (VR2)	_	.60	.56
Numerical (NR2)		_	.65
Abstract (AR2)			_

Table 14: Product-moment Correlations between the GRT2 Sub-scales (n = 1441)

THE STRUCTURE OF THE REASONING TESTS

Specifically we are concerned that the test's sub-scales are correlated with each other in a meaningful way. For example, we would expect the different sub-scales of a reasoning test to be moderately correlated as each will be measuring a different facet of general reasoning ability. Thus if such sub-scales are not correlated with each other we might wonder whether each is a good measure of reasoning ability. Moreover, we would expect different facets of verbal reasoning ability (e.g. vocabulary, similarities etc.) to be more highly correlated with each other than they are with a measure of numerical reasoning ability. Consequently, the first way in which we might assess the validity of a reasoning test is by exploring the relationship between the test's sub-scales.

CRTB	Verbal	Numerical
Verbal	1	.352
Numerical	.352	1

Table 15: Product-moment Correlations between CRTB Verbal & Numerical (n=352)

THE GRADUATE REASONING TESTS (GRT1)

Table 13, which presents Pearson Product-moment correlations between the three sub-scales of the GRT1 demonstrates two things. Firstly, the relatively strong correlations between each of the sub-scales indicate that each is measuring one facet of an underlying ability. This is clearly consistent with the design of this test, where each sub-scale was intended to assess a different facet of reasoning ability or mental alertness. Secondly, the fact that each sub-scale accounts for less than 30% (r < .55) of the variance in the other sub-scales indicates that the Verbal, Numerical and Abstract Reasoning sub-scales of the GRT1 are measuring different facets of reasoning ability, as they were designed to. Moreover, this is what we would in fact predict from research in the area intelligence testing (Heim, 1970).

THE GENERAL REASONING TESTS (GRT2)

Table 14, which presents Pearson Product-moment correlations between the three sub-scales of the GRT2 demonstrates two things. Firstly, the relatively strong correlations between each of the sub-scales indicate that each is measuring one underlying characteristic, which in this case we might assume to be reasoning ability or mental alertness. Thus these relatively strong correlations between the sub-scales are consistent with our intention to construct a test which measures general reasoning ability. Secondly, the fact that each sub-scale accounts for less than 45% (r < .65) of the variance in the other sub-scales indicates that the Verbal, Numerical and Abstract Reasoning sub-scales of the GRT2 are still measuring distinct aspects of reasoning ability.

THE CRITICAL REASONING TESTS (CRTB)

Table 15, which presents the Pearson Product moment correlation between the two sub-tests of the CRTB, demonstrates that while the Verbal and Numerical sub-tests are marginally correlated, they nevertheless measuring quite distinct abilities, sharing only 10% of common variance.

THE CONSTRUCT VALIDITY OF THE REASONING TESTS

As an evaluation of construct validity, the Psytech Reasoning Tests were administered with other widely used measures of similar constructs.

In the case of the General and Graduate reasoning tests, the AH series was considered to be a suitable external measure. The AH Series of tests is one of the most widely respected range of reasoning tests which have been developed on a U.K. population. Within this series there are tests which have been specifically designed for use on both the general population (AH2/ AH3/ AH4) and the graduate population (AH5/AH6). Developed by Alice Heim (1968, 1974) of Cambridge University the AH series has become something of a benchmark against which to compare the performance of other reasoning tests.

As the original Critical Thinking Appraisal, The Watson-Glaser (W-GCTA) (Watson & Glaser (1991) has set the standard in the assessment of abilities that are of relevance in management decision-making.

Thus we chose to explore the construct validity of the GRT1 and the GRT2 by comparing their performance against that of the AH3 and AH5 respectively, while the construct validity of the CRTB is examined by comparing its performance against both the AH5 and the Watson-Glaser Critical Thinking Appraisal.

THE RELATIONSHIP BETWEEN THE GRT1 AND AH5

Table 16 presents product-moment correlations between the sub-scales and the total scale scores of the AH5 and the GRT1. The AH5, which has been developed for use on a graduate population, has a Verbal/Numerical sub-scale which combines verbal and numerical items and an Abstract, or Diagrammatic, reasoning sub-scale. These along with the total scale score on the AH5 (the sum of the sub-scales) were correlated with the GRT1 sub-scale scores and the total scale score. The correlations with the total scale scores were included within this table, even though the total scale scores are simple composites of the sub-scale scores, as they provide a measure of general (g), rather than specific (e.g. verbal, numerical etc.) mental aptitudes.

Table 16 provides clear support for the concurrent validity of the GRT1 against the AH5. The correlations between each of the GRT1 sub-scales and their comparable AH5 sub-scales are high, indicating that they are measuring similar constructs. In addition, this table provides some evidence in support of the discriminant validity of these sub-scales. That is to say, each of the GRT1 sub-scales is more highly correlated with its comparable sub-scale on the AH5 than it is with the AH5 sub-scale measuring a different specific mental ability. For example, while the VR1 has a correlation of .69 with the Verbal/Numerical sub-scale of the AH5, its correlation with the Abstract sub-scale is only .35. In addition, the extremely high correlation (r=.84) between the total scale scores of the AH5 and the GRT1 indicates that, as a whole, this battery of tests is a good measure of general reasoning ability (g).

THE RELATIONSHIP BETWEEN THE GRT2 AND AH3

Table 17 presents Pearson Product-moment correlations (n=81) between each of the GRT2 sub-scales with each of the sub-scale scores of the AH3 and the total scale score. In addition to the high correlations between these sub-scales it is also worth noting the extremely high correlation (r=.82) between the total scale scores on these two tests. These results clearly demonstrate that the GRT2 is measuring the trait of general reasoning ability which is assessed by the AH3. We should however note that the correlations between the sub-scales provide no clear support for the discriminant validity of the GRT2. That is to say, the correlations between each of the GRT2 sub-scales and their respective AH3 sub-scales (e.g. the VR2 with the AH3 Verbal) are not significantly higher than are the correlations across sub-scales (e.g. the VR2 with the AH3 Numerical and Verbal).

SUB-SCALE	VR1	NR1	AR1	TOTAL
AH5 Verbal/Numerical	.69	.70	.35	.70
AH5 Abstract	.51	.67	.72	.74
AH5 Total	.69	.79	.65	.84

Table 16: Product-moment correlations between the GRT1 and AH5 Sub-scales

SUB-SCALE	VR2	NR2	AR2	GRT Total
AH5 Verbal	**.63**	.63	.61	.73
AH3 Numerical	.58	**.76**	.76	.70
AH5 Perceptual	.54	.55	**.56**	.76
AH5 Total	.70	.78	.78	**.82**

Table 17: Product-moment Correlations between the GRT2 and AH3 Sub-scales

RELATIONSHIP WITH GRT2 AND OTHER MEASURES

As a part of a number of data-collection exercises, the GRT2 was applied with a number of alternative measures, namely the technical Test Battery (TTB2) and Clerical Test Battery (CTB2).

Technical Test Battery (TTB2)

A sample of 94 trainee Mechanical apprentices completed both the GRT2 and the Technical Test Battery as part of a validation exercise. The GRT2 sub-scales register modest correlations with the components measures of the Technical Test Battery, although this is no more than would be expected from different aptitude measures with none exceeding .50 (see Table 18).

Clerical Test Battery (CTB2)

A sample of 54 clerical staff working for a major bank completed the Verbal reasoning Test (VR2) as part of an assessment of Clerical aptitudes which included components of the Clerical Test Battery (CTB2).

The strongest observed correlation was not with the Spelling measure (SP2) as expected but with Office Arithmetic (NA2). Examination of NA2 does reveal a fairly high verbal problem-solving element, which may explain this. The Clerical Checking Test (CC2) only registered a modest correlation which is as expected of a measure which relies only to a limited extent on general ability (see Table 19).

THE RELATIONSHIP BETWEEN THE CRTB AND AH5

In Tables 20 and 21 we present two sets of data supporting the concurrent validity of the VCR1 and NCR1. The first data set was collected trial versions of these two tests which contained each of the stem passages/tables which appear in the final tests along with approximately 80% of the final items. This data was originally collected as part of the test construction process in order to check that the trial items we had constructed were measuring reasoning ability, and not some other construct (e.g. reading ability, numerical checking etc.). This is particularly important when constructing critical reasoning tests as it is easy to construct items which are better measures of checking ability than they are of reasoning ability. That is to say, items which simply require carefully scanning and memorising the text in order to successfully complete them, rather than having to correctly draws logical inferences from the text. As can be seen from the table, our trial items clearly appear to be measuring ability, as scores on both the VCR1 and NCR1 are strongly correlated with the AH5.

Table 21 presents the correlations between the Verbal/Numerical sub-scale of the AH5 and the final version of the two critical reasoning tests. This data therefore provides evidence in support of the fact that the final versions of these two tests measure reasoning ability rather than some other construct (i.e. verbal or numerical checking ability). As was noted above, when developing critical reasoning tests it is particularly important to demonstrate that the tests are measuring reasoning ability, and not checking ability.

The above data clearly demonstrates, as does the previous data set, that both the VCR1 & NCR1 are measuring reasoning ability. The size of these correlations indicate that the Verbal/Numerical sub-scale of the AH5 and the two critical reasoning tests share no more than 35% of common variance, clearly demonstrating that these tests are measuring different, but related, constructs. This is what we would predict, given that the VCR1 & NCR1 were developed to measure critical reasoning, rather than be 'pure' measures of mental ability, or intelligence. Given the nature of critical reasoning, we would expect a candidate's scores on these tests not only to reflect general reasoning ability or intelligence (g), but also to reflect verbal and numerical comprehension, reading ability, reading speed and numerical ability and precision.

Technical Test Battery	**VR2**	**NR2**	**AR2**
MRT2	.45	.45	.38
SRT2	.35	.47	.46
VAC2	.34	.40	.40

Table 18: Pearson Correlations of GRT2 with Technical test Battery

CTB2 Sub-scales	**VR2**
Office Arithmetic (NA2)	.51
Clerical Checking (CC2)	.37
Spelling (SP2)	.34

Table 19: Pearson Correlations of GRT2 with Clerical Test Battery

CRTB	**AH5**
VRC1	.51
NRC1	.52

Table 20: Product-moment Correlations between the Experimental Versions of the VCR1 & NCR1 and the Verbal Numerical Sub-scale of the AH5

CRTB	**AH5**
VRC1	.60
NRC1	.51

Table 21: Product-moment Correlations between VCR1 & NCR1 and AH5 Verbal/Numerical

RELATIONSHIP BETWEEN THE CRTB AND WATSON-GLASER CRITICAL THINKING APPRAISAL

Table 22 details the relationship between the CRTB and the Watson-Glaser Critical Thinking Appraisal. The relationship between the total score on CRTB and W-GCTA is moderate, although this may be due to its absence of numerical content. However, unexpectedly, the CRTB Verbal sub-scale does not appear to have a higher correlation with the W-GCTA than the numerical. In summary, while CRTB does not appear to be measuring exactly the same construct as the W-GCTA, the domains do overlap to such an extent that this provides some evidence that the CRTB is a measure of critical thinking.

THE RELATIONSHIP BETWEEN THE CRTB AND THE MULTI-DIMENSIONAL APTITUDE BATTERY

The CRTB was correlated with Jackson's MAB to assess the relative positioning of the CRTB within a broad range of abilities. Table 23 details the relationship between each MAB sub-scale and the Verbal and Numerical components of the CRTB as well as the CRTB total score. The CRTB total correlates .60 with the MAB total and is equally related to both the MAB Verbal and Performance scales (.57 and .52). More specifically at the sub-scale level, the CRTB total, relates significantly to three of each of the Verbal and Performance sub-scales. It is more strongly related to the Information, Arithmetic and Object Perception sub-scales (.60, .58, .58). This it would appear that the CRTB total measures a composite of both Verbal and Performance abilities.

When the CRTB is divided into its sub-scales, the Verbal does not appear to be strongly related to the MAB Total but is related to the MAB verbal scales in particular, Information and Vocabulary.

As expected, the Numerical subscale relates more closely with the MAB Performance scale (.48) and also correlates significantly with MAB Arithmetic and Spatial as it does with the Verbal which can be expected by the higher verbal content in VCR1 than more classic measures of verbal ability.

	WGCTA
Verbal	.38
Numerical	.38
CRTB Total	.57

Table 22: Product-moment Correlations between VCR1 & NCR1 and WGCTA

	Verbal	Numerical	Total
MAB Total	.24	.48	.60
MAB Performance	.04	.48	.52
MAB Verbal	.43	.44	.57
Information	.29	.32	.60
Comprehension	.25	.44	.52
Arithmetic	.24	.45	.58
Similarities	.22	.33	.39
Vocabulary	.32	.27	.40
Digit Symbol	.09	.37	.39
Picture Completion	.14	.38	.44
Spatial	.19	.50	.58
Picture Arrangement	.15	.34	.42
Object Assembly	.09	.23	.44

Table 23: Product-moment Correlations between VCR1 & NCR1 and MAB (n=154)

CRITERION-RELATED VALIDITY

In this section, we provide details of number of studies in which the reasoning tests have been used as part of a pilot study on a sample of job incumbents on whom performance data was available.

INSURANCE SALES

A sample of 86 Telephone Sales Personnel with a leading Motor Insurance group completed the GRT2 as part of a validation study. The results of the GRT2 were correlated with a number of performance criteria as detailed in Table 24.

The pattern of correlations suggests that while there is a relationship between GRT2 and performance measures, this is not always in the anticipated direction. In fact some notable negative correlations were observed, indicating that for some performance criteria, higher reasoning ability may be less desirable. However, the strongest correlations were found between GRT2 sub-scales and 'Ins_Nce' and the VR2 with overall sales. The consistent negative correlations with 'Sales PH' would require further examination.

BANKING

A sample of 118 retail bankers completed the GRT2 and a personality measure OPP as part of a concurrent validation exercise. Participants were rated on a range of competencies which focused primarily on personal qualities as opposed to abilities. As expected, GRT2 failed to relate strongly to the overall competency rating which based on a composite of ratings covering such diverse areas as Orderliness, Planning, Organising, Teamwork etc. GRT2 did correlate with those ratings which were associated with skill areas namely, numerical and software related work (see Table 25).

SERVICE ENGINEERS

A leading International Crane & heavy lifting equipment servicing company tested a sample of 46 service engineers on the GRT2 battery and OPP. Their overall performance was rated by supervisors on a behaviourally anchored five point scale. From Table 26, the Verbal (VR2) is found to be strongly related to rated performance (r=.46) and the Abstract (AR2) relating moderately with the same.

PRINTERS

A major local newspaper group with the largest number of local titles in the United Kingdom sought to examine whether tests could predict the job performance of experienced printers. A sample of 70 completed the GRT2 battery as well as a number of other measures including the OPP (Occupational Personality Profile) and TTB (Technical Test Battery). Each of the group were assessed on a number of performance criteria by supervisors. In addition, test data were correlated with the results of a job sample print test which was administered at selection stage. Table 27 details the results of this study.

Some noteworthy correlations were registered with GRT2 and performance measures. Firstly, overall performance is highly correlated with the Abstract (AR2) but also moderately with the Verbal and Numerical sub-scales. The job sample criterion measure generally registers higher correlations with each of the GRT2 sub-scales, reaching .41 with the Abstract (AR2). Perhaps more surprising, the Abstract correlates .56 with a supervisor's rating of Initiative, consistent with the behavioural descriptions used for this performance rating, which point to not simply taking initiative but being able to successfully resolve problem situations.

The single totally objectively derived performance measure was time-keeping, which was based on an electronic time and attendance recording system. Only one significant correlation was observed with Abstract (AR2) although this was negative, suggesting that those with higher ability tended to have a poorer time and attendance record.

	Verbal	**Numerical**	**Abstract**
Sales PH	-.26	-.27	-.27
Conversions%	-.20	-.21	-.29
Ins_Nce	.32	.25	.33
Sales	.27	.14	.13

Table 24: Pearson Product Moment Correlations with GRT2 and job performance in an Insurance Setting

Sales PH	Sales per hour
Conversions %	Percent of Conversions from other policies
Ins_NCE	Composite Training Outcome Measure
Sales	Total Value of Policies sold

GRT2	**PERF_ARI**	**PERF_SOF**	**Competency**
VR2	.14	.03	-.02
NR2	**.29**	**.32**	.12
AR2	**.31**	**.28**	.01

Table 25: Pearson Product Moment Correlations with GRT2 and Performance criteria in Banking (N=118)

GRT2	**Overall**
Measure	**Performance**
Verbal	**.46**
Numerical	-.24
Abstract	**.28**

Table 26: Pearson Product Moment correlations between GRT2 & Service Engineer Performance

Criterion	**Verbal**	**Numerical**	**Abstract**
Overall Performance	.26	.28	.36
Performance Job Sample	.33	.30	.41
Initiative	.40	.44	.56
Time Keeping	_	_	.32

Table 27: Correlations Between GRT2 & Printer Performance Criteria (N=70)

FINANCIAL SALES CONSULTANTS

A sample of 100 trainee Financial Consultants from a major financial services group completed the GRT2 as part of a validation study. Table 28 details the correlations between GRT2 and end of year examinations.

The results indicate that the Numerical (NR2) appears to be the best predictor of examination performance with correlations (up to .46), although the Abstract (AR2) also registers some highly notable correlations (up to .44). Only the Verbal (VR2) fails to relate to any of the examination results which is perhaps somewhat unexpected.

TRAINING APPLICANTS FOR CAR COMPONENT TRAINING COURSE

A large training company used the General Reasoning Test to investigate the ability profiles of successful/non-successful applicants for training on a car components assembly task. The criterion of success used was only in part determined by training outcomes, with other factors also played a part, such as perceived attitude, temperament etc. A sample of 150 applicants was used for the study.

Both the Verbal (VR2) and Abstract (AR2) register moderate correlations with success on the programme. The numerical fails to relate with this criterion (see Table 29).

Criterion	**VR2**	**NR2**	**AR2**
Protection Clusters Result	.10	.31	.35
Pension exam Results	.04.	.40	.32
Seller Induction Exam Results	.18	.26	.32
Aggregate Result	.11	.46	.44
Financial Planning Certificate Result	.13	.44	.42

Table 28: Correlations between GRT2 & Proficiency Criteria

GRT2 Sub-scale	**Success**
Verbal	.27
Numerical	.16
Abstract	.30

Table 29: Correlations between OPP & Successful Applicant for Component Course

5 ADMINISTRATION INSTRUCTIONS

BEFORE STARTING THE QUESTIONNAIRE

Put candidates at their ease by giving information about yourself, the purpose of the questionnaire, the timetable for the day, if this is part of a wider assessment programme, and how the results will be used and who will have access to them. Ensure that you and other administrators have switched off mobile phones etc.

The instructions below should be read out verbatim and the same script should be followed each time the GRT1 is administered to one or more candidates. Instructions for the administrator are printed in ordinary type. Instructions designed to be read aloud to candidate incorporate a grey shaded background, italics and speech marks.

If this is the first or only questionnaire being administered give an introduction as per or similar to the following example (prepare an amendment if not administering all tests in this battery):

> "From now on, please do not talk among yourselves, but ask me if anything is not clear. Please ensure that any mobile telephones, pagers or other potential distractions are switched off completely. We shall be doing three tests: verbal, numerical and abstract reasoning. The tests take 8, 10 and 10 minutes respectively to complete. During the test I shall be checking to make sure that you are not making any accidental mistakes when filling in the answer sheet. I will not be checking your responses to see if you are answering correctly or not."

WARNING: It is most important that answer sheets do not go astray. They should be counted out at the beginning of the test and counted in again at the end.

Continue by using the instructions **EXACTLY** as given. Say:

DISTRIBUTE THE ANSWER SHEETS

Then ask:

> "Has everyone got two sharp pencils, an eraser, some rough paper and an answer sheet."

Rectify any omissions, then say:

> "Print your surname, first name and title clearly on the line provided and indicate your title, sex and age by ticking the appropriate boxes. Please insert today's date which is [ ] "

Walk around the room to check that the instructions are being followed.

WARNING: It is vitally important that test booklets do not go astray. They should be counted out at the beginning of the session and counted in again at the end.

DISTRIBUTE THE BOOKLETS WITH THE INSTRUCTION

> "Please do not open the booklet until instructed."

Remembering to read slowly and clearly, go to the front of the group and say:

> "Please open the booklet at Page 2 and follow the instructions for this test as I read them aloud."

Pause to allow booklets to be opened.

> "This test is designed to assess your understanding of words and relationships between words. Each question has six possible answers. One and only one is correct in each case. Mark your answer by filling in the number box on your answer sheet that corresponds to your choice. You now have a chance to complete the four example questions on Page 3 in order to make sure that you understand the test.
>
> Please attempt the example questions now, marking your answers in boxes E1 to E4."

Indicate section.

While the candidates are doing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody is looking at the actual test items during the example session. When all have finished (allow a maximum of two minutes) give the answers as follows:

> "The answer to Example 1 is number 2, sick means the same as ill.
> The answer to Example 2 is number 3, you drive a car and fly an aeroplane.
> The answer to Example 3 is number 5, wood is the odd one out..
> The answer to Example 4 is number 4, as both heavy and light have a relationship to weight.
>
> Is everyone clear about the examples?"

Then say:

> "REMEMBER:
>
> Time is short, so when you begin the timed test, work as quickly and as accurately as you can.
>
> If you want to change an answer, simply erase your first choice and fill in your new answer.
>
> There are a total of 35 questions and you have 8 minutes in which to answer them.
>
> If you reach the end before time is called you may review your answers if you wish.
>
> If you have any questions please ask now, as you will not be able to ask questions once the test has started."

Then say very clearly:

> "Is everyone clear about how to do this test?"

Deal with any questions appropriately, then, starting stop-watch or setting a count-down timer on the word **BEGIN** say:

> "Please turn over the page and begin."

Answer only questions relating to procedure at this stage, but enter in the Administrator's Test Record any other problems which occur. Walk around the room at appropriate intervals to check for potential problems.

At the end of the 8 minutes, say:

> "Stop now please and turn to Page 12."

NB: If this is the final test to be used in this battery, instead of the above line, please turn to the instructions in **Ending Test Session** on the final page of this section. If you are skipping a test, please find the appropriate test bookmark for your next test in the margin of the page and replace the above line as necessary.

You should intervene if candidates continue after this point.

Then say:

> "We are now ready to start the next test. Has everyone still got two sharpened pencils, an eraser, some unused rough paper?"

If not, rectify, then say:

> "The next test follows on the same answer sheet, please locate the section now."

Indicate section.

Check for understanding, then remembering to read slowly and clearly, go to the front of the group and say:

> "Please ensure that you are on Page 12 of the booklet and follow the instructions for this test as I read them aloud."

Pause to allow page to be found.

> "This test is designed to assess your ability to work with numbers. Each question has six possible answers. **One and only one** is correct in each case. Mark your answer by filling in the appropriate box that corresponds to your choice on the answer sheet.
>
> You now have a chance to complete the four example questions on Page 13 in order to make sure that you understand the test. Please attempt the example questions now, marking your answers in the example boxes.

Indicate section.

While the candidates are doing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody is looking at the actual test items during the example session. When all have finished (allow a maximum of two minutes) give the answers as follows:

> "The answer to Example 1 is number 5, the sequence goes up in twos.
> The answer to Example 2 is number 4, as all other fractions can be reduced further.
> The answer to Example 3 is number 2, 100 is 10 times 10.
> The answer to Example 4 is number 5, the journey will take 1 hour and 30 minutes.
>
> Is everyone clear about the examples?"

Then say:

> "Time is short, so when you begin the timed test work as quickly and as accurately as you can.
>
> If you want to change an answer, fully erase your first choice and fill in your new choice of answer.
>
> There are a total of 25 questions and you have 10 minutes in which to attempt them.
>
> If you reach the **end** before time is called you may review your answers to the numerical test if you wish, but do not go back to the verbal test.
>
> If you have any questions please ask now, as you will not be able to ask questions once the test has started."

Then say very clearly:

> "Is everyone clear about how to do this test?"

Deal with any questions, appropriately, then, starting stop-watch or setting a count-down timer on the word **BEGIN** say:

> "Please turn over the page and begin"

Answer only questions relating to procedure at this stage, but enter in the Administrator's Test Record any other problems which occur. Walk around the room at appropriate intervals to check for potential problems.

At the end of the 10 minutes, say:

> "Stop now please and turn to Page 20"

NB: If this is the final test to be used in this battery, instead of the above line, please turn to the instructions in **Ending Test Session** on the final page of this section. If you are skipping a test, please find the appropriate test book-mark for your next test in the margin of the page and replace the above line as necessary.

You should intervene if candidates continue after this point.

Then say:

> "We are now ready to start the next test. Has everyone still got two sharpened pencils, an eraser, some unused rough paper?"

If not, rectify, then say:

> "The next test follows on the same answer sheet, please locate the section now."

Indicate section.

Check for understanding, then remembering to read slowly and clearly, go to the front of the group and say:

> "Please ensure that you are on Page 20 of the booklet and follow the instructions for this test as I read them aloud."

Pause to allow page to be found.

> In this test you will have to work out the relationship between abstract shapes and patterns.
>
> Each question has six possible answers. One and only one is correct in each case. Mark your answer by filling in the appropriate box that corresponds to your chosen answer on your answer sheet. You now have a chance to complete the three example questions on Page 21 in order to make sure that you understand the test.
>
> Please attempt the example questions now, marking your answers in the example boxes."

Indicate section.

While the candidates are doing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody is looking at the actual test items during the example session. When all have finished, (allow a maximum of two minutes) give the answers as follows:

> "The answer to Example 1 is number 5, as the series alternates between 2 and 4 squares as does the direction of the two squares which return to their original position.
>
> The answer to Example 2 is number 4, as all of the other options have an open side to one of the boxes.
>
> The answer to Example 3 is number 6, as this is a mirror image of the pattern.
>
> Is everyone clear about the examples?"

Then say:

> "Time is short, so when you begin the timed test, work as quickly and as accurately as you can.
>
> If you want to change an answer, fully erase your first choice, and fill in your new choice of answer.
>
> There are a total of 25 questions and you have 10 minutes in which to attempt them.
>
> If you reach the end before time is called, you may review your answers to the abstract test, but do not go back to the previous tests.
>
> If you have any questions please ask now, as you will not be able to ask questions once the test has started."

Then say very clearly:

> "Is everyone clear about how to do this test?"

Deal with any questions appropriately, then, starting stop-watch or setting a count-down timer on the word **BEGIN** say:

> "Please turn over the page and begin."

Answer only questions relating to procedure at this stage, but enter in the Administrator's Test Record any other problems which occur. Walk around the room at appropriate intervals to check for potential problems.

At the end of the 10 minutes:

ENDING THE TEST SESSION

Say:

> "Stop now please and close your booklet'

You should intervene if candidates continue after this point.

COLLECT THE ANSWER SHEETS AND THE TEST BOOKLETS, ENSURING THAT ALL MATERIALS ARE RETURNED (COUNT BOOKLETS AND ANSWER SHEETS)

Then say:

> "Thank you for completing the Graduate Reasoning Test."

6

REFERENCES

Binet. A (1910) *Les idées modernes sur les enfants* Paris: E. Flammarion.

Cronbach L.J. (1960) *Essentials of Psychological Testing (2nd Edition)* New York: Harper.

Galton F. (1869) *Heriditary Genuis* London: MacMillan.

Gould, S.J. (1981). *The Mismeasure of Man.* Harmondsworth, Middlesex: Pelican.

Heim, A.H. (1970). *Intelligence and Personality*. Harmondsworth, Middlesex: Penguin.

Heim, A.H., Watt, K.P. and Simmonds, V. (1974). *AH2/AH3 Group Tests of General Reasoning; Manual.* Windsor: NFER Nelson.

Jackson D.N. (1987) *User's Manual for the Multidimensional Aptitude Battery* London, Ontario: Research Psychologists Press.

Johnson, C., Blinkhorn, S., Wood, R. and Hall, J. (1989). *Modern Occupational Skills Tests: User's Guide.* Windsor: NFER-Nelson.

Budd R.J. (1991) *Manual for the Clerical Test Battery:* Letchworth, Herts UK: Psytech International Limited

Budd R.J. (1993) *Manual for the Technical Test Battery*: Letchworth, Herts UK: Psytech International Limited

Stern W (1912) *Psychologische Methoden der Intelligenz-Prüfung.* Leipzig, Germany: Barth

Terman, L.M. et. al., (1917). *The Stanford Revision of the Binet-Simon scale for measuring intelligence.* Baltimore: Warwick and York.

Watson & Glaser (1980) *Manual for the Watson-Glaser Critical Thinking Appraisal* Harcourt Brace Jovanovich: New York

Yerkes, R.M. (1921). Psychological examining in the United States army. *Memoirs of the National Academy of Sciences*, 15.