PSYTECH INTERNATIONAL

General & Graduate Reasoning Tests

Published by PSYTECH INTERNATIONAL LIMITED
The Grange, Church Road,
Pulloxhill, Bedfordshire MK45 5HE
Tel 01525 720003
Fax 01525 720004
Email sales@psytech.com

1. THEORETICAL OVERVIEW

1.1 THE ROLE OF REASONING TESTS IN PERSONNEL SELECTION AND ASSESSMENT

Whilst much useful information can be gained from the standard (unstructured) job interview, the interview nonetheless suffers from a number of significant weaknesses. Perhaps the most important of these is that the interview is a very unreliable way to judge a person's aptitudes and abilities. This is because it is an unstandardised assessment procedure that does not directly allow one person's performance to be compared with another's performance.

Clearly, the interview can provide a useful opportunity to probe each applicant in depth about their work experience, and their understanding of the job/task requirements. In this regard it is noteworthy that structured interviews, which directly focus on the core competencies needed to successfully meet the job requirements, have been shown to have much greater validity than the standard unstructured job interview (Schmidt & Hunter, 1998). Such competency focussed interviews can be used to assess applicants' understanding of the core tasks and skills involved in a particular job. They can also provide an opportunity to present applicants with hypothetical work situations and explore their ability to explain the reasoning behind the decisions they would take in such situations. However competency focussed, structured interviews do not provide a reliable, standardised way to assess an applicant's ability to solve novel, complex problems that require the use of logic and reasoning ability.

Reasoning tests, on the other hand, do just this; providing a reliable, standardised way to assess an applicant's ability to use logic to solve complex problems. As such, reasoning tests are likely to have a significant role to play in many selection decisions. Thus it is not surprising that Schmidt & Hunter (1998), in their seminal review of the research literature, concluded that over 85 years of research has clearly demonstrated that tests of general mental (reasoning) ability have consistently been found to be the single best predictor of job performance.

From the perspective of assessing a respondent's reasoning ability, the unstandardised idiosyncratic nature of interviews makes it impossible to directly compare one applicant's ability with another's. Not only do interviews not provide an objective baseline against which to contrast interviewees' differing performances but, moreover, different interviewers typically come to radically different conclusions about the same applicant. Not only do applicants respond differently to different interviewers asking ostensibly the same questions, but what applicants say is often interpreted quite differently by different interviewers. In such cases we have to ask which interviewer has formed the 'correct' impression of the candidate, and to what extent any given interviewer's evaluation of the candidate reflects the interviewer's preconceptions and prejudices rather than reflecting the candidate's performance.

There are similar limitations on the range and usefulness of the information that can be gained from application forms or CV's. Whilst work experience and qualifications may be pre-requisites for many occupations, past performance does not necessarily predict whether a candidate will perform well or badly in a new position. Moreover, a person's educational and occupational achievements to date are likely to be limited by the opportunities they have had and, as such, may not reflect their true potential. Reasoning tests enable us to avoid many of these problems. Not only do they prove an objective measure of a person's ability, but also they assess the person's potential, and not just their achievements to date.

1.2 THE ORIGINS OF REASONING TESTS

The assessment of mental, or reasoning ability is one of the oldest areas of research interest in psychology. Gould (1981) has traced attempts to scientifically measure mental acuity, or intelligence, to the work of Sir Francis Galton in the late 1800s. Primarily interested in exploring the nature of genius, Galton (1869) assessed thousands of people using a range of primitive tests which mostly assessed reaction time, co-ordination and other motor skills. Prior to Galton's (1869) pioneering work, the assessment of mental ability had focussed on phrenologists' attempts to assess intelligence by measuring the size of people's heads!

Reasoning tests, in their present-day form, were first developed by Binet (1910); 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 he 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 answer each question. From the child's performance on this test it was possible to derive the child's mental age. If, for example, a child performed at the level of the average 10 year old on Binet's test, then that child was classified as having a mental age of 10, regardless of the child's chronological age.

The concept of the Intelligence Quotient (IQ) was developed by Stern (1912), from Binet's notion of mental age. Stern defined IQ as mental age divided by chronological age multiplied by 100. Previous to Stern's work, chronological age had been subtracted from mental age to provide a measure of mental alertness. Stern on the other hand showed that it was more appropriate to take the ratio of these two constructs, to provide a measure of the child's intellectual

development, that was independent of the child's age. 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 famous Stanford-Binet IQ test. IQ tests were first used for selection by the American military during the First World War, when Yerkes (1921) tested 1.75 million soldiers with the Army Alpha and Army Beta tests. Thus by the end of the war, the assessment of general mental ability had not only firmly established its place within the discipline of academic psychology, but had also demonstrated its utility for aiding the selection process.

1.3 THE CONCEPTS OF FLUID AND CRYSTALLISED INTELLIGENCE

The idea of general mental ability, or general intelligence, was first conceptualised by Spearman in 1903. He reflected on the popular notion that some people are more academically able than others, noting that people who tend to perform well in one intellectual domain (e.g. science) also tend to perform well in other domains (e.g. languages, mathematics, etc.). He concluded that an underlying factor termed general intelligence, or 'g', accounted for this tendency for people to perform well across a range of areas, while differences in a person's specific abilities or aptitudes accounted for their tendency to perform marginally better in one area than in another (e.g. to be marginally better at French than they are at Geography).

Spearman, in his 1904 paper, outlined the theoretical framework underpinning factor analysis; the statistical procedure that is used to identify the shared factor ('g') that accounts for a person's tendency to perform well (or badly) across a range of different tasks. Subsequent developments in the mathematics underpinning factor analysis, combined with advances in computing, meant that after the Second World War psychologists were able to begin exploring the structure of human mental abilities using these new statistical procedures.

Being most famous for his work on personality, and in particular the development of the 16PF, the pioneering work that Raymond B. Cattell (1967) did on the structure of reasoning abilities has often been overlooked. Through an extensive research programme, Cattell and his colleagues identified that 'g' (general intelligence) could be decomposed into two highly correlated subtypes of mental ability, which he termed fluid and crystallised intelligence.

Fluid intelligence is reasoning ability in its most abstract and purest form. It is the ability to analyse novel problems, identify the patterns and relationships that underpin these problems and extrapolate from these using logic. This ability is central to all logical problem solving and is crucial for solving scientific, technical and mathematical problems. Fluid intelligence tends to be relatively independent of a person's educational experience and, being the 'purest' form of reasoning ability, is often viewed as assessing a person's potential level of reasoning ability, independent of the cultural and educational experiences they have had to date. It is typically assessed by tests of abstract reasoning ability.

Crystallised intelligence, on the other hand, consists of fluid ability as it is evidenced in culturally valued activities. High levels of crystallised intelligence are evidenced in a person's good level of general knowledge, their extensive vocabulary and their ability to reason using words and numbers. In short, crystallised intelligence is the product of cultural and educational experience in interaction with fluid intelligence. Crystallised intelligence is often considered as assessing a person's current level of attained reasoning ability, rather than their potential ability, and is typically assessed by tests of verbal and numerical reasoning ability.

1.4 THE RELATIONSHIP BETWEEN REASONING ABILITY AND OCCUPATIONAL PERFORMANCE

From their review of over 85 years of research into the validity of different selection methods, Schmidt & Hunter (1998) concluded that reasoning tests have consistently been found to be the best predictors of job performance, with graphology (not surprisingly) having been found to be the least valid predictor of job performance. They also reported that in addition to predicting job performance, reasoning tests have consistently been found to predict the effectiveness of staff training programmes, with those staff who have higher levels of reasoning ability benefiting more from training than those of lower reasoning ability.

Using meta-analysis to aggregate results across different studies, Schmidt & Hunter (1998) found that reasoning tests have average validity coefficients of 0.51 for predicting job performance and of 0.56 for predicting trainability. Not surprisingly, they also found that reasoning tests were much more predictive of a person's performance in professional/managerial roles (with aggregate validities of 0.58) than they were predictive of a person's performance in unskilled jobs (with aggregate validities of 0.23), and that the inclusion of a personality test, alongside a reasoning test, further improves the prediction of job performance.

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

2.1 ITEM FORMAT

As noted above, general reasoning (mental) ability can be decomposed into a number of specific subtypes of ability or aptitude. Some of these subtypes of reasoning ability assess very specific aptitudes that are only modestly correlated with general reasoning ability. (Examples of very specific aptitudes, which are only modestly correlated with general reasoning ability, are mechanical and spatial reasoning ability and critical reasoning ability.) Research in the area of intelligence testing has repeatedly demonstrated that the three aptitude domains that are most consistently correlated with general reasoning ability are: verbal, numerical and abstract reasoning ability (Heim, 1970). Consequently the GRT1 and GRT2 were specifically developed to assess these domains of ability and comprise of three subtests, each of which assesses one of these subtypes of general reasoning ability.

Verbal and numerical reasoning ability are, as their respective names indicate, the ability to: use words and numbers in a rational way; correctly identify logical relationships between these entities and draw conclusions and inferences from them. Abstract reasoning assesses the ability to correctly identify the logical relationships between abstract patterns, shapes and geometric designs.

To ensure that the GRT1 and GRT2 assess reasoning ability in its broadest sense, rather than assessing extremely narrow, specific aptitudes, each subtest consists of a number of different item formats; each of which is known to be a reliable and valid multiple choice format for assessing reasoning ability (Heim, 1970). The GRT1 and GRT2 Verbal Reasoning subtests (VR1 and VR2 respectively) consist of items assessing: the multiple (possible) meanings of words; the ability to classify words into categories or groups; the ability to perceive the relationships between pairs of words. The GRT1 and GRT2 Numerical Reasoning subtests (NR1 and NR2 respectively) consist of items assessing: the ability to solve numerical problems using logic; the ability to classify numbers into categories or groups; the ability to perceive the relationships between pairs of numbers; the ability to understand number sequences and extrapolate the next term in the sequence. The GRT1 and GRT2 Abstract Reasoning subtests (AR1 and AR2 respectively) consist of items assessing: the ability to perceive the logical relationships between pairs of abstract geometric designs; the ability to understand the logical relationships that underpin a sequence of abstract designs, and extrapolate the next term in the sequence; the ability to classify abstract designs into categories or groups.

2.2 TEST CONSTRUCTION

Research has clearly demonstrated that in order to accurately assess reasoning ability it is necessary to use tests which have been specifically designed to measure the ability being assessed in the population on which the test is intended to be used. This ensures that the test is appropriate for the particular group being assessed. For example, a test designed for those of average ability will not accurately distinguish between people of high ability, as all the respondents' 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 respondents, with all their scores clustering towards the bottom of the scale, but also as the questions will mostly be too difficult for the respondents to answer, they are likely to lose motivation, thereby further reducing their scores.

Therefore both a graduate and a general population version of the reasoning were developed. These are termed the GRT1 and GRT2 respectively. The GRT2 was designed specifically to assess applicants for roles that require a general level of reasoning ability. These include general administrative, accounting and financial roles, as well as technician roles. The GRT1 was designed to assess applicants for graduate, managerial and professional roles, and other jobs that require an above average level of reasoning ability.

The initial item pool was trialled on students enrolled in tertiary and higher education, as well as on a sub-sample of respondents in full-time employment in a range of occupations. Following extensive trialling, a subset of items for each test, that had high levels of internal consistency (corrected item-whole correlations of 0.3 or greater), and of graded difficulty, were selected for inclusion in the GRT1 and GRT2 subtest.

3. THE PSYCHOMETRIC PROPERTIES OF THE GRT1 and GRT2

3.1 STANDARDISATION

Normative data allows us to compare an individual's score on a standardised scale against the typical score obtained from a clearly defined group of respondents (e.g. graduates, the general population, etc.).

To enable any respondent's score on the GRT1 and GRT2 to be meaningfully interpreted, these tests were standardised against populations similar to that on which they have been designed to be used (e.g. people in technical, managerial, professional and scientific roles). Such standardisation ensures that the scores obtained on these tests can be interpreted by relating them to a relevant distribution of scores.

3.2 RELIABILITY

The reliability of a test assesses the extent to which the variation in test scores is due to true differences between people, on the characteristic being measured – in this case fluid intelligence – or to random measurement error.

Reliability is generally assessed using one of two different methods; one assesses the stability of the test's scores over time, the other assesses the internal consistency, or homogeneity, of the test's items.

3.2.1 RELIABILITY: TEMPORAL STABILITY

Also known as test-retest reliability, this method for assessing a test's reliability involves determining the extent to which a group of people obtain similar scores on the test when it is administered at two points in time. In the case of reasoning tests, where the ability being assessed does not change substantially over time (unlike personality), the two occasions when the test is administered may be many months apart. If the test were perfectly reliable, that is to say test scores were not influenced by any random error, then respondents would obtain the same score on each occasion, as their level of intelligence would not have changed between the two points in time when they completed the test. In this way, the extent to which respondents' scores are unstable over time can be used to estimate the test's reliability.

Stability coefficients provide an important indicator of a test's likely usefulness. If these coefficients are low, then this suggests that the test is not a reliable measure and is therefore of little practical use for assessment and selection purposes.

3.2.2 RELIABILITY: INTERNAL CONSISTENCY

Also known as item homogeneity, this method for assessing a test's reliability involves determining the extent to which, if people score well on one item they also score well on the other test items. If each of the test's items were a perfect measure of intelligence, that is to say the score the person obtained on the items was not influenced by any random error, then the only factor that would determine whether a person was able to answer each item correctly would be the item's difficulty. As a result, each person would be expected to answer all the easier test items correctly, up until the point at which the items became too difficult for them to answer. In this way, the extent to which respondents' scores on each item are correlated with their scores on the other test items, can be used to estimate the test's reliability.

The most commonly used internal consistency measure of reliability is Cronbach's (1960) alpha coefficient. If the items on a scale have high inter-correlations with each other, then the test is said to have a high level of internal consistency (reliability) and the alpha coefficient will be high. Thus a high alpha coefficient indicates that the test's items are all measuring the same thing, and are not greatly influenced by random measurement error. A low alpha coefficient on the other hand suggests that either the scale's items are measuring different attributes, or that the test's scores are affected by significant random error. If the alpha coefficient is low, this indicates that the test is not a reliable measure, and is therefore of little practical use for assessment and selection purposes.

3.3 VALIDITY

The fact that a test is reliable only means that the test is consistently measuring a construct, it does not indicate what construct the test is consistently measuring. The concept of validity addresses this issue. As Kline (1993) notes 'a test is said to be valid if it measures what it claims to measure'.

An important point to note is that a test's reliability sets an upper bound for its validity. That is to say, a test cannot be more valid than it is reliable because if it is not consistently measuring a construct, it cannot be consistently measuring the construct it was developed to assess. Therefore, when evaluating the psychometric properties of a test, its reliability is usually assessed before addressing the question of its validity. There are two principle ways in which a test can be said to be valid.

3.3.1 VALIDITY: CONSTRUCT VALIDITY

Construct validity assesses whether the characteristic which a test is measuring is psychologically meaningful and consistent with how that construct is defined. Typically, the construct validity of a test is assessed by demonstrating that the test's results correlate other major tests which measure similar constructs and do not correlate with tests that measure different constructs. (This is sometimes referred to as a test's convergent and discriminant validity). Thus

demonstrating that a test which measures fluid intelligence, is more strongly correlated with an alternative measure of fluid intelligence than it is with a measure of crystallised intelligence, would be evidence of the measure's construct validity.

3.3.2 VALIDITY: CRITERION VALIDITY

This method for assessing the validity of a test involves demonstrating that the test meaningfully predicts some real-world criterion. For example, a valid test of fluid intelligence would be expected to predict academic performance, particularly in science and mathematics.

Moreover, there are two types of criterion validity - predictive validity and concurrent validity. 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 of fluid intelligence predict future GCSE maths results? Concurrent validity assesses whether the scores on a test can be used to predict a criterion which is available at the time the test was completed, e.g. can a test of fluid intelligence predict a scientist's current publication record?

3.4 GRT1/2 STANDARDISATION

The GRT1 (Graduate/Professional Adults norm) was standardised on a sample of 1,270 British adults, drawn from a variety of managerial, professional and graduate occupations. The mean age of the standardisation sample was 30.1 years (age range 19-58), with 34% of the sample being women. 79% of the sample identified themselves as being of White European ethnic origin; 7% of Indian origin; 6% of Pakistani origin; 1% of Black (i.e. Black Caribbean, Black African origin or Black Other origin) and 6% of Other (i.e. Bangladeshi, Chinese, etc.) ethnic origin.

The GRT2 (General Working Age Adults norm) was standardised on a sample of 6,618 British adults of working age, drawn from a variety of occupations, including; call centre and production line staff, and staff in sales, clerical, administrative and accounts roles. The mean age of the standardisation sample was 34.7 years (age range 17-67), with 35% of the sample being women. 91% of the sample identified themselves as being of White European ethnic origin; 3% of Indian origin; 3% of Pakistani origin; 1% of Black (i.e. Black Caribbean, Black African or Black Other) origin and 3% of Other (i.e. Bangladeshi, Chinese, etc.) ethnic origin.

For the GRT1 and GRT2, a variety of other international norms (e.g. Australian professionals, South African adults) and norms for specific groups (e.g. accountants, etc.), are available on the GeneSys assessment software system.)

3.5 GRT1/2 RELIABILITY: INTERNAL CONSISTENCY

Tables 1 and 2 present alpha coefficients for the GRT1 and GRT2 on a number of different (English speaking) international samples, and samples of respondents drawn from different occupational groups. Inspection of Tables 1 and 2 indicates that these alpha coefficients are above 0.8 for most samples, indicating that these tests have good levels of internal consistency reliability across a range of respondents of different nationalities and from different occupations.

3.6 GRT1/2 RELAIBILITY: TEST-RETEST

As noted above, test-retest reliability estimates the test's reliability by assessing the temporal stability of the test's scores. As such, test-retest reliability provides an alternative measure of reliability to internal consistency estimates of reliability, such as the alpha coefficient. Table 3 reports test-retest reliability statistics for the GRT subscales. The test-retest reliability of the GRT1 was calculated from a sample of 71 undergraduates who completed the test on two occasions, three months apart. The test-retest reliability of the GRT2 was calculated from a sample of 54 college students who completed the test on two occasions, two weeks apart. Inspection of these data indicates that the GRT1 and GRT2 display good levels of temporal stability (test-rest reliability).

Table 1 – Alpha coefficients for the GRT1 subscales

Subscale	Alpha						
VR1	.81	.82	.95	.81	.81	.78	.80
NR1	.84	.84	.95	.86	.84	.80	.81
AR1	.71	.86	.85	.72	.73	.72	.75
	UK Professionals (n=1,270)	Australian Professionals (n=1,662)	NZ Professionals (n=2,776)	Programmers/ Systems Analysts (n=121)	Accountants (n=152)	Civil Servants (n=322)	Solicitors/ Paralegals (n=106)

Table 2 – Alpha coefficients for the GRT2 subscales

Subscale	Alpha							
VR2	.78	.81	.83	.89	.79	.83	.81	.86
NR2	.86	.88	.88	.87	.82	.89	.83	.83
AR2	.81	.80	.82	.85	.79	.82	.80	.78
	UK Adults (n=6,618)	Australian Adults (n=11,020)	NZ Adults (n=5,183)	South African Adults (n=666)	Telesales Staff (n=112)	Civil Servants (senior grades) (n=118)	Therapists/ Care Staff (n=105)	Call Centre Staff (n=123)

Table 3 – Test-Retest reliability coefficients for the GRT1/2 subscales

Subscale	GRT1	GRT2
Verbal	.79	.81
Numerical	.81	.84
Abstract	.78	.78
	Undergrads (n=71)	Students (n=54)

3.7 GRT1/2 CONSTRUCT VALIDITY

3.7.1 THE RELATIONSHIP BETWEEN THE GRT SUBSCALES

Tables 4 and 5 present (respectively) the correlations between the GRT1 and GRT2 subscales. These data demonstrate that the GRT subscales are significantly correlated with each other, as would be expected given that each of these subscales is assessing a different facet of general reasoning ability. Most significantly, however, these correlations are not sufficiently large as to suggest that each of these subscales is assessing the same construct. These data therefore provide strong support for both the convergent and discriminant construct validity of the GRT subscales.

Table 4 – Correlations between the GRT1 subscales (n=2,776)

	VR1	NR1	AR1
VR1	—		
NR1	.44	—	
AR1	.41	.46	—

Table 5 – Correlations between the GRT2 subscales (n=5,183)

	VR2	NR2	AR2
VR2	—		
NR2	.62	—	
AR2	.59	.63	—

3.7.2 THE RELATIONSHIP BETWEEN THE GRT1/2 AND THE AH3/5

The AH series of tests are a well respected set of tests that were designed in the 1960s and 70s to assess reasoning ability in both the graduate (AH5 and AH6) and general (AH2, AH3 and AH4) population. Developed by Alice Heim (1968, 1974) and her colleagues at Cambridge University, these tests continue to be widely used in occupational assessment and selection.

Tables 6 and 7 present (respectively) the correlations between the GRT1 and AH5 subscales, and between the GRT2 and AH3 subscales, on graduate (n=) and general (n=) populations samples that completed these tests for experimental purposes. These correlations are all substantial in size, providing strong support for convergent construct validity of the GRT1 and GRT2.

Table 6 – Correlations between the GRT1 and AH5 subscales (n=52)

AH5 Subscale	VR1	NR1	AR1
Verbal - Numerical	.69	.70	.35
Abstract	.51	.67	.72

Table 7 – Correlations between the GRT2 and AH3 subscales (n=46)

AH3 Subscale	VR2	NR2	AR2
Verbal	.63	.63	.58
Numerical	.61	.76	.76
Perceptual	.54	.55	.65

3.7.3 THE RELATIONSHIP BETWEEN THE GRT2 AND THE TTB2

A sample of 94 mechanical engineering apprentices completed the GRT2 and the TTB2 (Technical Test Battery). The TTB2 contains subsets assessing; mechanical reasoning (MRT2), spatial reasoning (SRT2) and visual acuity (VAT2). The correlations between the GRT2 and TTB2 subscales are presented in Table 8.

As would be expected, while all these correlations were statistically significant ($p < .01$) they were mostly relatively modest in size, reflecting the fact that the TTB2 assesses fairly specific technical aptitudes rather than general reasoning ability. When taken in combination with the correlations between the GRT2 and the AH3 (reported above), these results provide support for both the convergent and discriminant construct validity of the GRT2.

Table 8 – Correlations between the GRT2 and TTB2 subscales (n=94)

	VR2	NR2	AR2
MRT2	.45	.45	.38
SRT2	.35	.47	.46
VAT2	.34	.40	.40

3.7.4 THE RELATIONSHIP BETWEEN THE GRT2 AND THE CTB2

A sample of 54 clerical staff working for a major British bank completed the Verbal Reasoning (VR2) subscale of the GRT2 along with Clerical Test Battery (CTB2) as part of an assessment of their clerical skills. The CTB2 contains subscales which assess spelling (SP2), basic (office) arithmetic (NA2) and clerical checking (CC2). As would be predicted, the verbal subscale of the GRT2 was only modestly correlated with spelling ($r=.34$, $p<.05$) and clerical checking ($r=.37$, $p<.05$) ability, with these being very specific aptitudes that are only modestly correlated with general reasoning ability. The substantial correlation that was observed between the verbal subscale of the GRT2 and arithmetical ability ($r=.51$, $p<.01$) probably reflects the fact that the (NA2) not only requires respondents to perform basic arithmetic calculations, but also to solve a variety of numerical problems that require the use of logic and an understanding of a variety of basic numerical concepts (e.g. percentages, etc.).

3.7.5 MEAN GRT SUBSCALE SCORES BY OCCUPATIONAL GROUP

Tables 9 and 10 present (respectively) the means scores (with associated 95% confidence intervals) on the GRT1 and GRT2 subscales, broken down by occupational group. As would be expected, different occupational groups show mean differences in reasoning ability, which clearly reflect the differing competencies required for these different occupations. The 95% confidence intervals indicate that the differences in reasoning ability, between many of these occupational groups, are not due to chance effects. These significant differences in mean GRT subscale scores, between different occupational groups, therefore provides further support for the construct validity of the GRT1 and GRT2.

Table 9 – Mean GRT1 subscale scores (with associated 5% confidence intervals) by occupational group

	Professional/ Managerial	IT Professionals	Administration	Sales & Marketing	Customer Service
VR2	16.2 ±0.6	15.3 ±1.2	13.8 ±1.4	13.1 ±1.2	12.1 ±1.2
NR2	13.0 ±0.6	13.7 ±1.1	12.6 ±1.2	10.9 ±1.7	10.6 ±1.3
AR2	13.5 ±0.4	13.6 ±0.7	13.7 ±0.9	12.2 ±0.8	12.1 ±1.0
	n=335	n=91	n=62	n=70	N=59

Table 10 – Mean GRT2 subscale scores (with associated 5% confidence intervals) by occupational group

	Managerial	Customer Service	Sales	Clerical	Manual (skilled)	Semi-skilled/ Unskilled
VR2	23.5 ±1.0	22.3 ±0.6	22.3 ±1.0	19.6 ±1.6	17.5 ±1.5	16.1 ±2.9
NR2	15.3 ±1.1	13.6 ±0.6	13.4 ±0.9	11.1 ±1.6	12.5 ±1.6	9.9 ±2.8
AR2	16.8 ±0.9	15.6 ±0.6	15.5 ±0.7	13.4 ±1.4	13.7 ±1.3	13.9 ±2.8
	n=120	n=288	n=173	n=78	n=70	n=17

3.7.6 MEAN GRT SUBSCALE SCORES BY EDUCATIONAL LEVEL

Tables 11 and 12 present (respectively) the means scores (with associated 95% confidence intervals) on the GRT1 and GRT2 subscales, broken down by the respondents' educational level. As would be expected, mean reasoning ability varies by educational level. The 95% confidence intervals indicate that the differences in mean GRT subscale scores, between respondents of differing educational levels, are not due to chance effects. These significant differences in mean GRT subscale scores, by educational level, provide strong support for the construct validity of the GRT1 and GRT2.

3.7.7 THE RELATIONSHIP BETWEEN THE GRT AND 15FQ+ INTELLECTANCE

Intellectance is a meta-cognitive variable that assesses a person's perception of their general mentality ability. Whilst it is a personality factor, rather than an ability factor, intellectance has nonetheless consistently been found to correlate with objective assessments of reasoning ability. As such it would be expected to be modestly correlated with the GRT subscales. The correlations between the GRT subscale scores and intellectance was examined on a sample of respondents who had completed the 15FQ+ and either the GRT1 or GRT2 as part of an assessment and selection process. These correlations are presented in Table 13. While modest in size, all these correlations are statistically significant, thereby providing further support for the concurrent construct validity of the GRT1 and GRT2.

Table 11 – Mean GRT1 subscale scores (with associated 5% confidence intervals) by educational level

	Postgraduate qualification(s)	University degree	Tertiary Education	Completed Secondary Education
VR2	16.5 ±1.0	15.9 ±0.5	13.4 ±1.8	12.7 ±1.2
NR2	13.2 ±1.0	12.8 ±0.5	11.5 ±1.4	9.7 ±1.1
AR2	13.3 ±0.8	13.9 ±0.4	13.1 ±0.9	11.7 ±0.9
	n=92	N=416	n=60	n=79

Table 12 – Mean GRT2 subscale scores (with associated 5% confidence intervals) by educational level

	Postgraduate Qualification(s)	University Degree	Tertiary Education	Industry/Trade Qualification(s)	Completed Secondary Education	Not Completed Secondary Education
VR2	23.8 ±1.7	23.0 ±0.8	20.9 ±1.1	19.9 ±1.3	20.0 ±0.8	15.0 ±2.3
NR2	15.7 ±1.6	15.3 ±0.8	13.3 ±1.1	12.2 ±1.3	12.6 ±0.7	9.1 ±2.5
AR2	16.8 ±1.5	17.3 ±0.7	15.3 ±1.1	13.6 ±1.2	15.5 ±0.7	12.6 ±2.0
	n=66	n=222	n=101	n=69	n=247	n=27

Table 13 – Correlations between the GRT1 and GRT2 subscale scores and Intellectance (15FQ+ β)

	VR1	NR1	AR1	VR2	NR2	AR2
R	.38‡	.27†	.21*	.32†	.28†	.22*
Sample size	n=128	n=128	n=128	n=98	n=98	n=98

*p<.05 †p<.01 ‡p<.001

3.7.8 THE RELATIONSHIP BETWEEN THE GRT1 NUMERICAL REASONING (NR1) TEST AND THE ART

The Abstract Reasoning Test (ART) is a measure of fluid intelligence that has been designed for use with graduate populations. It is similar in format to the Ravens Advanced Progressive Matrices (RAPM) test. A sample of 209 graduate level bankers completed the Numerical Reasoning (NR1) subscale of the GRT1 along with the ART, as part of an assessment exercise. These two tests were found to be significantly correlated ($r=.49$, $p<.001$) with each other, providing good support for the concurrent construct validity of the NR1.

3.7.9 THE RELATIONSHIP BETWEEN THE GRT2 AND THE CRTB2

A sample of 25 undergraduates completed (for experimental purposes) the verbal and numerical subscales of the GRT2 along with the verbal and numerical subscales of the Critical Reasoning Test Battery - 2nd Edition (CRTB2). The verbal subscales of the GRT2 and the CRTB2 were found to be substantially correlated with each other ($r=.57$, $p<.001$), as were the numerical subscales of the GRT2 and CRTB2 ($r=.51$, $p<.001$). These large correlations therefore provide strong support for the concurrent construct validity of the GRT2.

3.8 GRT1/2 CRITERION VALIDITY

3.8.1 PREDICTING THE PERFORMANCE OF PRINTERS

A sample of 70 printers, working for a major local newspaper group, completed the GRT2 (and a personality measure) as part of the selection process, along with a work sample test. Their performance was subsequently rated (by their line managers) on a number of job relevant criteria (e.g. time keeping, initiative, etc,), which were combined to form an overall job performance rating. The correlations between the GRT2 subscale scores and these performance criteria are reported in Table 14. These correlations demonstrate that the GRT2 subscales are predictive of performance, and thus provide support for the predictive criterion validity of the GRT2.

Table 14 – Correlations between the GRT2 sub-scales and performance in a sample (n=70) printers

	Job sample test	Overall performance
VR2	.33†	.26*
NR2	.30*	.28*
AR2	.41‡	.36‡

*p<.05 †p<.01 ‡p<.001

3.8.2 PREDICTING THE PERFORMANCE OF RETAIL BANKERS

A sample of 118 retail bankers completed the GRT2 (and a personality measure) as part of a concurrent test validation exercise. Participants were rated for their overall competency, as well as being rated for their numerical skills/accuracy and IT skills. The correlations between the GRT2 subscales and rated competency are presented in Table 16. The GRT2 was found to be significantly correlated with both numerical and IT skills, but not with overall performance. The failure of the GRT2 subscales to correlate with overall performance probably reflects the breadth of this composite performance criterion, which covered such diverse areas as: orderliness; planning; organising; team-working, etc.

Table 15 – Correlations between the GRT2 subscales and performance in a sample (n=118) of bankers

	Numerical skills	IT skills	Overall performance
VR2	.14	.03	-.02
NR2	.29*	.32†	.12
AR2	.31†	.28*	.01

*p<.01 †p<.01

3.8.3 PREDICTING PERFORMANCE IN FINANCIAL SERVICES EXAMINATIONS

A sample of 100 sales consultants in the (UK) financial services industry completed the GRT2 prior to enrolling on a training programme to prepare staff for financial services examinations. At the end of the course, their scores on the GRT2 were correlated with their examination results. These data are presented in Table 16. Inspection of this table indicates that both the numerical and abstract subscales of the GRT2 were significantly correlated with financial services examination performance. These results provide strong evidence of the predictive criterion validity of this test.

Table 16 – Correlations between the GRT2 subscales and the performance of a sample (n=100) of sales consultants on financial services examinations

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

*p<.01 †p<.01

3.8.4 PREDICTING SUCCESS ON A CAR COMPONENTS TRAINING COURSE

150 trainees on a car components training course completed the GRT2 as part of a test validation exercise. The trainees were classified as either being successful or unsuccessful on the basis of both the level of skill they acquired during the course and their work attitude and behaviour. The verbal (r=.27 p<.001) and abstract (r=.30 p<.001), but not the numerical (r=.16 n.s.), subscales of the GRT2 were found to be correlated with training success. Whilst modest in size, these statistically significant correlations nonetheless provide further evidence of both the ability of the GRT2 to predict training success, and thus provide further support for its concurrent criterion validity.

3.8.5 PREDICTING THE PERFORMANCE OF CONSULTANTS AND MANAGERS

A combined sample of managers (n=48) and consultants (n=20) working in the live stock industry in New Zealand completed the GRT2 (along with a number of personality measures) as part of a test validation exercise. Their performance was rated (by their line manager) on a number of criteria (listed in Table 17), which were also combined to form a global performance criterion. The GRT2 subscales were found to predict global performance, as well as predicting a number of the specific performance criterion. The correlations between the GRT2 subscales and the performance criteria are presented in Table 17. Most significantly, the GRT2 subscales were found to be most strongly correlated with those performance criteria that are related to mental ability (e.g. decision making, etc.), and less strongly correlated with those performance criteria that are related to personality (e.g. resilience, etc.) These correlations therefore demonstrate both convergent and discrimant validity, and provide strong support for the concurrent criterion validity of the GRT2.

3.8.6 PREDICTING MBA GRADES

A sample of 142 South African MBA students completed the GRT1 prior to commencing their course. The scores they obtained on a variety of management development options they were studying as part of their MBA were correlated with the GRT1 subscales. These correlations are reported in Table 18. (The sample size varies between courses depending upon the number of students enrolled in each course.)

Inspection of Table 18 indicates that the GRT1 scores were substantially correlated with the grades students obtained on a number of management courses. The fact that many of these correlations failed to research statistical significance, despite the magnitude of many of these effects, reflects the relatively small size of some of these samples. Due to the magnitude of many of these correlations these results do, nonetheless, provide strong support for the predictive criterion validity of the GRT1.

Table 17 – Correlations between the GRT2 subscales and the listed performance criteria in sample (n=68) of managers and consultants in the live stock industry

	VR2	NR2	AR2
Analytical Ability	.28*	.21	.41‡
Energy	.18	-.05	.02
Decision Making	.52‡	.36†	.25*
Resilience	.19	.05	.11
Interpersonal Skills	.29*	.08	.12
Planning and Organising	.22	-.10	.10
Persuasiveness	.30*	.21	.21
Technical Expertise	.44‡	.20	.37†
Creativity	.28*	.13	.12
Overall Job Performance	.32†	.20	.30*

*p<.05 †p<.01 ‡p<.001

Table 18 – Correlations between the GRT1 subscales and the grades obtained by a sample of MBA students on each of the listed courses

	VR2	NR2	AR2
Quantitative Management Techniques (n=21)	.41	.88‡	.33
Economics (n=8)	.58	.63	.56
Principles of Management Accounting (n=24)	.45*	.32	.30
Principles of Marketing Management (n=72)	.03	-.07	.30*
Human Resources Management (n=11)	.23	.46	.43
Financial Management (n=11)	.44	.45	.53
Operations Management(n=59)	.07	.12	.26*
Management Skills - I (n=16)	.27	.52*	.28
Management Skills - II (n=22)	.38	.54‡	.18

*p<.05 †p<.01 ‡p<.001

3.8.7 PREDICTING THE PERFORMANCE OF TRAINING ADVISORS

A sample (n=39) of advisors, working for a New Zealand industry and training advisory service, completed the GRT2 as part of a test validation exercise. Their performance was rated (by their line manager) on a number of competency dimensions, which were summed to form a composite (overall) measure of competency. The verbal subscale of the GRT2 was found to predict the overall rated performance of job

incumbents. However, neither the numerical (r=.15 n.s.) nor abstract (r=-.11 n.s.) subscales of the GRT2 were found to be related to overall job performance ratings.

3.9 GRT1/2 BIAS

Differences in mean scores on reasoning tests between different groups (i.e. differences in mean scores by gender, ethnicity, social class, etc.) have repeatedly been observed. Such mean group differences in scores can be attributed to two possible factors. Firstly, they may reflect real differences between populations in the characteristic(s) the test measures. (This is termed differential test impact.) Secondly, these group differences may reflect aspects of test bias. That is to say, they may be due to the test's items functioning differently between different groups. (This is termed differential item functioning - DIF.)

The issue of test bias and its assessment has rightly received considerable attention over the last decade. However, while a number of different methods have been developed for assessing DIF (see Camilli & Sheppard, 1994; Holland & Wainer, 1993), there is as yet no agreement as to which of the many methodologies that have been proposed is the best. As logistic regression is the methodology that is currently most widely used to assess DIF, this methodology was adopted to assess the presence (or absence) of uniform bias in the GRT items. The logic underpinning this methodology is as follows. If each item does not show uniform bias across groups, then it would be expected that the binary group effect variable (e.g. gender, ethnicity, etc.) would not predict each item's score once each person's level of reasoning ability has been controlled for, by entering their corrected total test score (i.e. their test score minus the score on the item which is being examined for DIF). That is to say, if an item does not display uniform bias, the *only* factor that should predict a person's success on that item is their level of reasoning ability, and not their group membership (i.e. gender, ethnicity, etc.).

Bias in the GRT2 items was examined on two samples; a sample of 279 British men and 252 British women, and a sample of 319 British adults of White European ethnic origin and a sample of 331 British adult's of varied (non-white) ethnic origin. The mixed ethnic sample consisted of the following ethnic groups: Indian (33%); Pakistani (28%); Black Caribbean (15%); Black African (8%); Black Other (5%); Chinese (5%); Bangladeshi (4%). (These different ethnic groups were combined to for one mixed ethnic sample in order to provide a sufficiently large sample for the tests for DIF to be sensitive to relatively small bias effects.) Tables 19, 20 and 21 present the logistic regression (maximum likelihood estimation) coefficients, and the associated significance level, for the group effects (ethnicity and gender) for the items on the verbal, numerical and abstract subscales of the GRT2 respectively.

Table 19 - Item bias statistics (British adults) for the GRT2 verbal subscale

Item No.	Mixed Ethnic vs. White European	Male vs. Female
1	ß=.102 p=.765	ß=.329 p=.153
2	ß=.227 p=.322	ß=.174 p=.501
3	ß=.205 p=.381	ß=.166 p=.540
4	ß=.044 p=.814	ß=.461 p=.030
5	ß=.315 p=.205	ß=.166 p=.557
6	ß=.443 p=.103	ß=.055 p=.850
7	ß=.300 p=.123	ß=.439 p=.043
8	ß=.011 p=.965	ß=.110 p=.675
9	ß=.109 p=.588	ß=.209 p=.373
10	ß=.191 p=.292	ß=.189 p=.566
11	ß=.543 p=.027	ß=.022 p=.916
12	ß=.251 p=.342	ß=.300 p=.182
13	ß=.202 p=.351	ß=.203 p=.388
14	ß=.190 p=.320	ß=.089 p=.665
15	ß=.308 p=.091	ß=.399 p=.087
16	ß=.561 p=.018	ß=.105 p=.629
17	ß=.241 p=.381	ß=.284 p=.205
18	ß=.091 p=.609	ß=.240 p=.232
19	ß=.151 p=.328	ß=.029 p=.901
20	ß=.249 p=.211	ß=.206 p=.380
21	ß=.351 p=.106	ß=.006 p=.981
22	ß=.179 p=.401	ß=.262 p=.301
23	ß=.128 p=.539	ß=.011 p=.964
24	ß=.247 p=.221	ß=.184 p=.349
25	ß=.272 p=.161	ß=.260 p=.228
26	ß=.233 p=.246	ß=.058 p=.796
27	ß=.004 p=.993	ß=.340 p=.090
28	ß=.093 p=.603	ß=.069 p=.722
29	ß=.195 p=.349	ß=.477 p=.040
30	ß=.220 p=.291	ß=.195 p=.262
31	ß=.399 p=.081	ß=.179 p=.216
32	ß=.104 p=.646	ß=.087 p=.711
33	ß=.262 p=.165	ß=.061 p=.429
34	ß=.217 p=.244	ß=.258 p=.231
35	ß=.091 p=.696	ß=.249 p=.290
	n=650	n=531

Table 20 – Item bias statistics (British adults) for the GRT2 numerical subscale

Item No.	Mixed Ethnic vs. White European	Male vs. Female
1	ß=.052 p=.713	ß=.095 p=.641
2	ß=.546 p=.023	ß=.291 p=.211
3	ß=.244 p=.170	ß=.150 p=.454
4	ß=.073 p=.734	ß=.276 p=.278
5	ß=.186 p=.452	ß=.112 p=.607
6	ß=.236 p=.258	ß=.136 p=.552
7	ß=.119 p=.589	ß=.019 p=.941
8	ß=.259 p=.260	ß=.271 p=.288
9	ß=.293 p=.116	ß=.432 p=.043
10	ß=.023 p=.918	ß=.194 p=.557
11	ß=.057 p=.791	ß=.123 p=.620
12	ß=.021 p=.919	ß=.029 p=.905
13	ß=.230 p=.256	ß=.192 p=.398
14	ß=.015 p=.940	ß=.409 p=.072
15	ß=.139 p=.464	ß=.396 p=.063
16	ß=.201 p=.343	ß=.287 p=.208
17	ß=.209 p=.217	ß=.044 p=.840
18	ß=.222 p=.229	ß=.030 p=.892
19	ß=.121 p=.527	ß=.291 p=.182
20	ß=.614 p=.014	ß=.464 p=.125
21	ß=.097 p=.478	ß=.163 p=.498
22	ß=.170 p=.582	ß=.256 p=.308
23	ß=.131 p=.541	ß=.130 p=.593
24	ß=.153 p=.460	ß=.395 p=.099
25	ß=.272 p=.174	ß=.094 p=.676
	n=650	n=531

Table 21 – Item bias statistics (British adults) for the GRT2 abstract subscale

Item No.	Mixed Ethnic vs. White European	Male vs. Female
1	ß=.043 p=.870	ß=.075 p=.715
2	ß=.045 p=.818	ß=.492 p=.027
3	ß=.421 p=.025	ß=.132 p=.505
4	ß=.156 p=.607	ß=.009 p=.969
5	ß=.101 p=.959	ß=.030 p=.894
6	ß=.210 p=.645	ß=.029 p=.881
7	ß=.267 p=.553	ß=.302 p=.385
8	ß=.239 p=.810	ß=.267 p=.200
9	ß=.129 p=.818	ß=.018 p=.933
10	ß=.058 p=.919	ß=.177 p=.483
11	ß=.218 p=.399	ß=.447 p=.034
12	ß=.141 p=.692	ß=.197 p=.390
13	ß=.140 p=.890	ß=.266 p=.333
14	ß=.202 p=.840	ß=.069 p=.732
15	ß=.226 p=.831	ß=.283 p=.429
16	ß=.186 p=.525	ß=.270 p=.367
17	ß=.188 p=.623	ß=.055 p=.801
18	ß=.164 p=.893	ß=.366 p=.078
19	ß=.297 p=.597	ß=.101 p=.647
20	ß=.083 p=.830	ß=.345 p=.089
21	ß=.435 p=.053	ß=.281 p=.230
22	ß=.105 p=.994	ß=.162 p=.498
23	ß=.292 p=.497	ß=.479 p=.023
24	ß=.304 p=.059	ß=.047 p=.823
25	ß=.379 p=.468	ß=.038 p=.127
	n=650	n=531

Inspection of Tables 19, 20 and 21 indicates that the GRT2 items show little bias by sex or ethnicity in British samples. While a few group effects are statistically significant (and a number approach statistical significance) these effects are likely to have occurred by chance. This is suggested by the observation that if the Bonferonni correction had been used to adjust significance levels for the number of multiple comparisons that have been made for each subscale, in order to avoid accepting the null hypothesis at the 5% level for any of the GRT2 verbal subscale items, a significance level of less than 0.001% would have had to have been adopted.

Ethnic bias in the GRT1 items was examined on two international samples of 408 respondents, one of which consisted of respondents from a variety of different ethnic backgrounds (living in Britain, Australia and New Zealand), the second consisted sample of respondents of White European ethnic origin matched for country of residence. Sex bias in the GRT1 was examined on two international samples of 401, one of which consisted of men (living in Britain, Australia and New Zealand), and the other consisted of women matched for country of residence. Tables 22, 23 and 24 present the logistic regression (maximum likelihood estimation) coefficients, and the associated significance level, for the group effects (ethnicity and

Table 22 – Item bias statistics for the GRT1 verbal subscale

Item No.	Mixed Ethnic vs. White European	Male vs. Female
1	ß=.361 p=.570	ß=.079 p=.684
2	ß=.012 p=.962	ß=.087 p=.640
3	ß=.076 p=.660	ß=.209 p=.280
4	ß=.094 p=.511	ß=.177 p=.205
5	ß=.086 p=.287	ß=.021 p=.925
6	ß=.236 p=.124	ß=.101 p=.606
7	ß=.148 p=.344	ß=.188 p=.413
8	ß=.292 p=.096	ß=.082 p=.664
9	ß=.117 p=.157	ß=.313 p=.118
10	ß=.038 p=.812	ß=.183 p=.301
11	ß=.352 p=.039	ß=.229 p=.221
12	ß=.051 p=.528	ß=.064 p=.727
13	ß=.162 p=.298	ß=.157 p=.395
14	ß=.181 p=.217	ß=.214 p=.279
15	ß=.145 p=.150	ß=.167 p=.212
16	ß=.044 p=.735	ß=.139 p=.428
17	ß=.078 p=.467	ß=.181 p=.352
18	ß=.044 p=.752	ß=.194 p=.273
19	ß=.262 p=.122	ß=.265 p=.193
20	ß=.240 p=.142	ß=.385 p=.054
21	ß=.048 p=.278	ß=.259 p=.213
22	ß=.124 p=.461	ß=.022 p=.909
23	ß=.346 p=.072	ß=.474 p=.035
24	ß=.341 p=.051	ß=.186 p=.359
25	ß=.108 p=.459	ß=.442 p=.022
26	ß=.063 p=.727	ß=.185 p=.345
27	ß=.347 p=.194	ß=.350 p=.078
28	ß=.113 p=.587	ß=.312 p=.206
29	ß=.208 p=.122	ß=.152 p=.532
30	ß=.018 p=.941	ß=.168 p=.355

Table 23 – Item bias statistics for the GRT1 numerical subscale

Item No.	Mixed Ethnic vs. White European	Male vs. Female
1	ß=.193 p=.729	ß=.306 p=.199
2	ß=.018 p=.918	ß=.264 p=.231
3	ß=.066 p=.721	ß=.427 p=.042
4	ß=.160 p=.666	ß=.070 p=.721
5	ß=.094 p=.614	ß=.132 p=.562
6	ß=.298 p=.052	ß=.027 p=.882
7	ß=.054 p=.767	ß=.002 p=.996
8	ß=.251 p=.104	ß=.111 p=.548
9	ß=.335 p=.031	ß=.239 p=.198
10	ß=.033 p=.834	ß=.341 p=.073
11	ß=.271 p=.082	ß=.349 p=.079
12	ß=.029 p=.860	ß=.196 p=.345
13	ß=.112 p=.481	ß=.070 p=.735
14	ß=.135 p=.310	ß=.206 p=.314
15	ß=.168 p=.616	ß=.095 p=.621
16	ß=.216 p=.174	ß=.200 p=.342
17	ß=.054 p=.744	ß=.153 p=.510
18	ß=.167 p=.354	ß=.067 p=.747
19	ß=.020 p=.904	ß=.231 p=.293
20	ß=.063 p=.745	ß=.181 p=.443
21	ß=.238 p=.083	ß=.130 p=.644
22	ß=.223 p=.091	ß=.323 p=.200
23	ß=.299 p=.074	ß=.009 p=.976
24	ß=.077 p=.787	ß=.563 p=.037
25	ß=.045 p=.894	ß=.326 p=.351

Table 24 – Item bias statistics for the GRT1 abstract subscale

Item No.	Mixed Ethnic vs. White European	Male vs. Female
1	ß=.151 p=.117	ß=.039 p=.818
2	ß=.202 p=.446	ß=.016 p=.966
3	ß=.145 p=.105	ß=.325 p=.237
4	ß=.149 p=.222	ß=.157 p=.476
5	ß=.256 p=.116	ß=.109 p=.598
6	ß=.024 p=.887	ß=.446 p=.024
7	ß=.169 p=.403	ß=.211 p=.086
8	ß=.033 p=.393	ß=.039 p=.813
9	ß=.299 p=.057	ß=.309 p=.081
10	ß=.042 p=.367	ß=.127 p=.456
11	ß=.278 p=.089	ß=.048 p=.794
12	ß=.146 p=.202	ß=.230 p=.180
13	ß=.090 p=.547	ß=.185 p=.234
14	ß=.149 p=.120	ß=.182 p=.176
15	ß=.008 p=.952	ß=.179 p=.166
16	ß=.022 p=.391	ß=.315 p=.056
17	ß=.159 p=.278	ß=.337 p=.067
18	ß=.127 p=.404	ß=.283 p=.098
19	ß=.036 p=.333	ß=.141 p=.446
20	ß=.156 p=.276	ß=.229 p=.228
21	ß=.345 p=.056	ß=.066 p=.725
22	ß=.051 p=.678	ß=.184 p=.374
23	ß=.094 p=.216	ß=.281 p=.141
24	ß=.109 p=.632	ß=.225 p=.325
25	ß=.424 p=.083	ß=.198 p=.432

Inspection of Tables 22, 23 and 24 indicates that the GRT1 items show little bias by sex or ethnicity in mixed international samples. While a few group effects are statistically significant (and a number approach statistical significance) these effects are likely to have occurred by chance. This is suggested by the observation that if the Bonferonni correction had been used to adjust significance levels for the number of multiple comparisons that have been made for each subscale, in order to avoid accepting the null hypothesis at the 5% level for any of the GRT2 verbal subscale items, a significance level in the order of 0.001% would have had to have been adopted.

APPENDIX I – GRT1 ADMINISTRATION INSTRUCTIONS

BEFORE STARTING THE QUESTIONNAIRE:-

Put candidates at their ease by giving information about: yourself; the purpose of the test; the timetable for the day; whether or not the questionnaire is being completed as 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 requested that all mobile phones have been switched off, etc.

The instructions below should be read out verbatim. The 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 candidates have lines marked above and below them, are in italics and are enclosed by speech marks.

If this is the first or only questionnaire being administered, give an introduction as per or similar to the following example:

"From now on please do not talk amongst yourselves, but ask me if anything is not clear. Please ensure that any mobile telephones, pagers or other potential distractions are switched off. We shall be doing three tests: a verbal, a numerical and an abstract reasoning test. 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 essential that answer sheets do not go astray. They should be counted out at the beginning of the session and counted in again at the end.

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 last name and first name clearly on the lines provided. Indicate your preferred title by checking the title box, then note your gender, age and ethnic origin. Please insert today's date which is [ ] in the space provided."

If biographical information is required, ask respondents to complete the biodata section. If answer sheets are to be scanned, explain and demonstrate how the ovals are to be completed, emphasising the importance of fully blackening the oval.

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 booklets until instructed to do so."

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 the relationships between words. Each question has six possible answers. One and only one is correct in each case. Mark your answers, by filling in the appropriate box that corresponds to your chosen answer, on your answer sheet."

Check that everyone has understood the instructions so far, then say:-

---

"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 E3."

Indicate the appropriate section on the answer sheet.

While the candidates are completing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody looks through the actual test items after completing the example questions. When everyone has finished the example questions (allow a maximum of one and half 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?"

Answer any questions, then say:-

"Before you begin the timed test, please note the following points:-

- *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 30 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 everybody clear about how to do this test?"

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

"Please turn over the page and begin."

Only answer questions relating to the test procedure at this stage, and enter in the Administrator's Test Record any problems which occur. Walk around the room at appropriate intervals to check for potential problems.

At the end of the 8 minutes say clearly:-

"STOP NOW please and turn to Page 12."

You should intervene if any candidates continue beyond this point.

NB: If this is the final test to be used in this battery, instead of the above line, please turn to the instructions at the bottom of Page 18 of this manual, detailing how to end the test session. If you are skipping the numerical subtest, bookmark the appropriate section of the administration instructions so you can easily find the point at which you should continue reading the protocol aloud.

Then say:-

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

If not, rectify this, then say:-

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

Indicate the appropriate section on the answer sheet.

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 understand numbers and the relationships between numbers. Each question has six possible answers. One and only one is correct in each case. Mark your answers by filling in the appropriate box, that corresponds to your chosen answer, on your answer sheet."

Check that everyone has understood the instructions so far, then say:-

"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 no."

Indicate the appropriate section on answer sheet.

While the candidates are completing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody looks through the actual test items after completing the example questions. When everyone has finished the example questions (allow a maximum of one and half 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?"

Answer any questions, then say:-

"Before you begin the timed test, please note the following points:-

- *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 25 questions and you have 10 minutes in which to answert 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 everybody clear about how to do this test?"

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

"Please turn over the page and begin."

Only answer questions relating to the test procedure at this stage, and enter in the Administrator's Test Record any problems which occur. Walk around the room at appropriate intervals to check for potential problems.

At the end of the 10 minutes say clearly:-

"STOP NOW please and turn to Page 20."

You should intervene if any candidates continue beyond this point.

NB: If this is the final test to be used in this battery, instead of the above line, please turn to the instructions at the bottom of Page 18 of this manual, detailing how to end the test session.

Then say:-

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

If not, rectify this, then say:-

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

Indicate the appropriate section on the answer sheet.

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.

"This test is designed to assess your ability to perceive and understand the relationships 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."

Check that everyone has understood the instructions so far, then say:-

"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."

Indicate the appropriate section on answer sheet.

While the candidates are completing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody looks through the actual test items after completing the example questions. When everyone has finished the example questions (allow a maximum of one and half 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?"

Answer any questions, then say:-

"Before you begin the timed test, please note the following points:-

- *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 answer.*
- *There are a total of 25 questions and you have 10 minutes in which to answer 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 verbal or numerical test.*
- *If you have any questions please ask now, as you will not be able to ask questions once the test has started."*

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

"Please turn over the page and begin."

Only answer questions relating to procedure at this stage, and 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 clearly:-

"STOP NOW please and close your booklet."

You should intervene if any candidates continue beyond this point.

COLLECT ANSWER SHEETS AND TEST BOOKLETS, ENSURING THAT ALL MATERIALS ARE RETURNED.

……

COUNT BOOKLETS AND ANSWER SHEETS AS YOU COLLECT THEM IN ORDER TO ENSURE THAT NONE HAVE GONE ASTRAY.

Then say:-

"Thank you for completing the Graduate Reasoning Test."

APPENDIX – GRT2 ADMINISTRATION INSTRUCTIONS

BEFORE STARTING THE QUESTIONNAIRE:-

Put candidates at their ease by giving information about: yourself; the purpose of the test; the timetable for the day; whether or not the questionnaire is being completed as 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 requested that all mobile phones have been switched off, etc.

The instructions below should be read out verbatim. The script should be followed each time the GRT2 is administered to one or more candidates. Instructions for the administrator are printed in ordinary type. Instructions designed to be read aloud to candidates have lines marked above and below them, are in italics and are enclosed by speech marks.

If this is the first or only questionnaire being administered, give an introduction as per or similar to the following example:

"From now on please do not talk amongst yourselves, but ask me if anything is not clear. Please ensure that any mobile telephones, pagers or other potential distractions are switched off. We shall be doing three tests: a verbal, a numerical and an abstract reasoning test. 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 essential that answer sheets do not go astray. They should be counted out at the beginning of the session and counted in again at the end.

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 last name and first name clearly on the lines provided. Indicate your preferred title by checking the title box, then note your gender, age and ethnic origin. Please insert today's date which is [ ] in the space provided."

If biographical information is required, ask respondents to complete the biodata section. If answer sheets are to be scanned, explain and demonstrate how the ovals are to be completed, emphasising the importance of fully blackening the oval.

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 booklets until instructed to do so."

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 the relationships between words. Each question has six possible answers. One and only one is correct in each case. Mark your answers, by filling in the appropriate box that corresponds to your chosen answer, on your answer sheet."

Check that everyone has understood the instructions so far, then say:-

---

"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 E3."

Indicate the appropriate section on the answer sheet.

While the candidates are completing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody looks through the actual test items after completing the example questions. When everyone has finished the example questions (allow a maximum of one and half 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 5, as dark means the opposite of light.

Is everyone clear about the examples?"

Answer any questions, then say:-

"Before you begin the timed test, please note the following points:-

- *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 everybody clear about how to do this test?"

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

---

"Please turn over the page and begin."

Only answer questions relating to the test procedure at this stage, and enter in the Administrator's Test Record any problems which occur. Walk around the room at appropriate intervals to check for potential problems.

At the end of the 8 minutes say clearly:-

"STOP NOW please and turn to Page 12."

You should intervene if any candidates continue beyond this point.

NB: If this is the final test to be used in this battery, instead of the above line, please turn to the instructions at the bottom of Page 23 of this manual, detailing how to end the test session. If you are skipping the numerical subtest, bookmark the appropriate section of the administration instructions so you can easily find the point at which you should continue reading the protocol aloud.

Then say:-

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

If not, rectify this, then say:-

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

Indicate the appropriate section on the answer sheet.

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 understand numbers and the relationships between numbers. Each question has six possible answers. One and only one is correct in each case. Mark your answers by filling in the appropriate box that corresponds to your chosen answer, on your answer sheet."

Check that everyone has understood the instructions so far, then say:-

"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 the appropriate section on answer sheet.

While the candidates are completing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody looks through the actual test items after completing the example questions. When everyone has finished the example questions (allow a maximum of one and half 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?"

Answer any questions, then say:-

"Before you begin the timed test, please note the following points:-

- *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 25 questions and you have 10 minutes in which to answer 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 everybody clear about how to do this test?"

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

"Please turn over the page and begin."

Only answer questions relating to the test procedure at this stage, and enter in the Administrator's Test Record any problems which occur. Walk around the room at appropriate intervals to check for potential problems.

At the end of the 10 minutes say clearly:-

"STOP NOW please and turn to Page 20."

You should intervene if any candidates continue beyond this point.

NB: If this is the final test to be used in this battery, instead of the above line, please turn to the instructions at the bottom of Page 23 of this manual, detailing how to end the test session.

Then say:-

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

If not, rectify this, then say:-

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

Indicate the appropriate section on the answer sheet.

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.

"This test is designed to assess your ability to perceive and understand the relationships 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".

Check that everyone has understood the instructions so far, then say:-

"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."

Indicate the appropriate section on the answer sheet.

While the candidates are completing the examples, walk around the room to check that everyone is clear about how to fill in the answer sheet. Make sure that nobody looks through the actual test items after completing the example questions. When everyone has finished the example questions (allow a maximum of one and half 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?"

Answer any questions, then say:-

"Before you begin the timed test, please note the following points:-

- *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 answer.*
- *There are a total of 25 questions and you have 10 minutes in which to answer 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 verbal or numerical test.*
- *If you have any questions please ask now, as you will not be able to ask questions once the test has started."*

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

"Please turn over the page and begin."

Only answer questions relating to procedure at this stage, and 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 clearly:-

"STOP NOW please and close your booklet."

You should intervene if any candidates continue beyond this point.

COLLECT ANSWER SHEETS AND TEST BOOKLETS, ENSURING THAT ALL MATERIALS ARE RETURNED.

……

COUNT BOOKLETS AND ANSWER SHEETS AS YOU COLLECT THEM IN ORDER TO ENSURE THAT NONE HAVE GONE ASTRAY.

Then say:-

"Thank you for completing the General Reasoning Test."

REFERENCES

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

Camilli, G. & Shepard, L. (1994). *Methods for identifying biased test items.* Thousands Oaks, CA: Sage.

Cattell, R. B. (1967). The theory of fluid and crystallised intelligence. *British Journal of Educational Psychology, 37,* 209-223.

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

Galton F. (1869). *Hereditary Genius.* London: MacMillan.

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

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

Holland, B. S. & Wainer, H. (1993). *Differential item functioning.* Hillside, NJ: Lawrence Erlbaum.

Kline, P. (1993). *Personality: The Psychometric View.* London: Routledge.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-273.

Spearman, C. (1904). General-intelligence; objectively defined determined and measured. *American Journal of Psychology, 15,* 210-292.

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

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

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