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

CONTENTS

1. THEORETICAL OVERVIEW

1.1 THE ROLE OF INTELLIGENCE TESTS IN PERSONNEL SELECTION AND ASSESSMENT

While much useful information can be gained from the standard job interview, the interview nonetheless suffers from a number of serious weaknesses. Perhaps the most important of these is that the interview has been shown to be 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 assess ability, but rather assesses a person's social skills and reported past achievements and performance.

Clearly, the interview provides a useful opportunity to probe each applicant in depth about their work experience, and explore how they present themselves in a formal social setting. Moreover, behavioural interviews can be used to assess an applicant's ability to 'think on their feet' and explain the reasoning behind their decision making processes. Assessment centres can provide further useful information on an applicant by assessing their performance on a variety of work-based simulation exercises. However, interviews and assessment centre exercises do not provide a reliable, standardised way to assess an applicant's ability to critically appraise written (textual) information and numerical data, and draw logical inferences from such information.

Critical reasoning tests, on the other hand, do just this; providing a reliable, standardised way to assess an applicant's ability to understand reports, tables and graphs and logically deduce the implications of such information. Thus, tests of critical reasoning ability are likely to play a significant role in the selection process. Most significantly, in this regard, Schmidt & Hunter (1998), in their seminal review of the research literature, note that over 85 years of research has clearly demonstrated that general mental (i.e. reasoning) ability is the single best predictor of job performance.

From the perspective of assessing a respondent's critical 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 base-line 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 prerequisites for certain occupations, in and of themselves they do not predict whether a candidate is likely to perform well or badly in a new position. Moreover, a person's educational and occupational achievements are likely to be limited by the opportunities they have had, and as such may not reflect their true potential. Reasoning tests, on the other hand, enable us to avoid many of these problems, by proving an objective measure of a person's ability, and by assessing their potential, rather than just their achievements to date.

1.2 THE ORIGINS OF INTELLIGENCE TESTS

The assessment of general mental ability, or intelligence, is one of the oldest areas of research interest in psychology. Gould (1981) has traced attempts to scientifically measure mental acuity, or ability, to the work of Galton in the late 1800s. Prior to Galton's (1869) pioneering research, 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 1904. 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 human intelligence 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 has been shown to be strongly determined by genetic factors. As such it is often considered to be the 'purest' form of intelligence, or 'innate mental ability', and is typically assessed by abstract reasoning tests.

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. As such it is assessed by traditional tests of verbal and numerical reasoning ability.

1.4 THE RELATIONSHIP BETWEEN CRITICAL REASONING AND INTELLIGENCE

The concept of critical thinking is usually traced back to the work of Dewey (1933), with Glaser (1941) having been the first person to have operastionalised this concept and to have produced the first measure of critical thinking. It is generally accepted that the ability to critically appraise written information and draw logical conclusions from this information (verbal critical reasoning ability) is more relevant to many jobs than is fluid intelligence (i.e. abstract reasoning ability), or general intelligence ('g'). As such, critical reasoning ability would be expected to be a better predictor of performance in many job roles than either general or fluid intelligence. While the theory of general intelligence predicts that people who have higher levels of 'g' will generally have higher levels of critical reasoning ability, the relationship between 'g' and critical reasoning ability would be expected to be at best modest. This is because it is only the general, or shared component of people's performance on reasoning test items that contributes to 'g', with the specific components of different test formats assessing specific types of reasoning ability. As such, critical reasoning can be considered to be a specific (job relevant) ability that is a facet (i.e. component) of crystallised intelligence.

Glaser's (1941) original conceptualisation of critical reasoning ability focussed solely on the ability to critically appraise written text. (Hence the Watson-Glaser Critical Thinking Appraisal - Watson & Glaser (1980) - only assesses verbal critical reasoning ability.) However, subsequent research has demonstrated that the ability to critically appraise numerical data, tables and graphs can be considered to be a different facet of critical reasoning ability that is relevant to a significant number of jobs roles. As such, more recent conceptualisations of critical reasoning ability have distinguished between verbal and numerical critical reasoning. Thus, in line with current thinking, the CRTB2 consists of verbal and numerical subtests that can either be administered individually or jointly,

for those roles which require incumbents to critically appraise and evaluate either textual and/or numerical information.

1.5 THE RELATIONSHIP BETWEEN GENERAL MENTAL 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 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 improved the prediction of job performance.

2. THE CRITICAL REASONING TEST BATTERY - 2ND EDITION

2.1 ITEM FORMAT

In constructing the items for the Verbal (VCR2) and Numerical (NCR2) Critical Reasoning tests a number of guide lines were followed, in order to ensure that the tests would be valid measures of critical reasoning ability. Most importantly, care was taken when writing the items to ensure that in order to correctly answer each question it was necessary to draw logical conclusions and inferences from the stem passage/table, and that it was not possible to answer an item correctly simply by checking the question against the information contained in the stem passage/table. This was done to ensure that the test was assessing critical (i.e. logical/deductive) reasoning, rather than simple verbal/numerical checking ability.

In order to achieve this goal for the Verbal Critical Reasoning (VCR2) test two further specific guidelines were set for the stem passages. 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 greatly affected by their 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 that was counter-intuitive, and was thus likely to create confusion, and thereby increase the importance of attention to detail (i.e. checking ability as opposed to critical reasoning ability) in obtaining the correct answer. Finally, to increase the acceptability of the test to respondents the themes for the stem passages were chosen to be relevant to a wide range of business situations. As a consequence, the stem passages used in the VCR2 were similar in many ways to the short articles found in the financial pages of a daily newspaper, or trade magazine.

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 quickly to lose motivation when completing the test, thereby further reducing their scores.

The two subtests of the CRTB2 were specifically designed to assess the Verbal Critical Reasoning (VCR2) and Numerical Critical Reasoning (NCR2) ability of people in scientific, engineering, financial and professional roles, as well as those who have to take strategic business decisions on the basis of written information or numerical data. As such the test items were developed for respondents of average, or above average, ability.

The initial item pool was trialled on undergraduate students as well as on a sub-sample of respondents in full-time employment in a range of professional, managerial and technical occupations. Following extensive trialling, a subset of items that had high levels of internal consistency (corrected item-whole correlations of .4 or greater), and were of graded difficult, were selected for inclusion in the VCR2 and NCR2.

2.3 REVISIONS FOR THE 2ND EDITION OF THE CRITICAL REASONING TEST BATTERY

The second edition of the Verbal and Numerical Critical Reasoning Test Battery has been revised to meet the following goals:

- To improve the face validity of the test items, thus increasing the test's acceptability to respondents.
- To modernise the items to reflect contemporary business and financial issues.
- To improve the tests' reliability and validity while maintaining the tests' brevity – with the CRBT2 being administrable in under an hour.
- To simplify test scoring.
- To make available hand scored as well as a computer scored version of the tests.
- To make an optional correction for guessing available for the VCR.

Perhaps the most significant change in the second edition of the VCR has been the incorporation of a correction for guessing. This obviates the problem that, due to the three-point response scale that is used in the verbal critical reasoning test, it is theoretically possible for respondents to get 33% of the items correct simply by guessing. While a variety of methods have been proposed for solving this potential problem (including the use of negative or harsh scoring criteria) we believe that a correction for guessing is the most elegant and practical solution to this issue.

The correction for guessing is based on the number of items the respondent gets wrong on the test. We know that to get these items wrong the respondent

must have incorrectly guessed the answer to these items. We can further assume that, by chance, the respondent incorrectly guessed the answer 66% of the time and correctly guessed the answer 33% of the time. Thus it is possible to estimate the number of correct guesses the respondent made from the number of incorrect responses they made. This correction can then be subtracted from the total obtained test score to adjust for the number of items the respondent is likely to have correctly guessed.

The use of this optional correction for guessing improves the test's score distribution and increases its power to discriminate between the respondents' 'true' ability levels. However, as the norm tables for corrected and uncorrected scores are significantly different from each other, the correction for guessing can only be used with those norm groups for which a corrected norm has been calculated; currently, only the general adult norm. Thus, while we believe that there are many advantages to using the correction for guessing, many test users may not wish to use the correction as a result of them wishing to use a norm other than the general professional adult norm.

When using the GeneSys assessment software, users wishing to correct VCR2 test scores for guessing must therefore use the "corrected" norm group. When hand scoring the Verbal Critical Reasoning test (hand scoring instruction are provide in Appendix II), it is important that users ensure that the correct norm table is used to standardise the scores on the VCR2; as the norms for the corrected and uncorrected scores are significantly different from each other. Thus, the user must take care to ensure that if they have **not** corrected the VCR2 scores for guessing they norm these scores using the norm table for the uncorrected scores (Appendix IV - Table 2), and if they have corrected the scores for guessing they norm these scores using the norm table for corrected scores (Appendix IV - Table 3).

3. THE PSYCHOMETRIC PROPERTIES OF THE CRTB2

3.1 STANDARDISATION

Normative data allows us to compare an individual's score on a standardised scale against the typical score obtained from a defined group of respondents (e.g. graduates, the general population, etc.). To enable any respondent's scores on the VCR2 and NCR2 to be meaningfully interpreted, the test was standardised against a population similar to that on which it has been designed to be used (e.g. people in technical, managerial, professional and scientific roles). Such standardisation ensures that the scores obtained on the CRTB2 subtests 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 verbal and numerical critical reasoning ability – 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 reasoning ability 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 critical reasoning ability, 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 intercorrelations 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 is consistent with how that construct is defined. Typically, the construct validity of a test is assessed by demonstrating that the test's results correlate with 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 construct validity). Thus demonstrating that a verbal critical reasoning test is more strongly correlated with an alternative measure of verbal critical reasoning than it is with a measure of fluid 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 numerical critical reasoning ability would be expected to predict success in finance and accountancy roles.

Moreover, there are two types of criterion validity- predictive validity and concurrent validity. Predictive criterion 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 numerical critical reasoning ability predict future accountancy examination results. Concurrent criterion 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 verbal critical reasoning ability predict an academic historian's current publication record.

3.4 CRTB2: STANDARDISATION

The CRTB2 was standardised on a sample of 4,625 adults of working age, drawn from a variety of professional, managerial and graduate occupations (Professional/Managerial norm). The mean age of the standardisation sample was 37.1 years (age range 19-66), with 30.1% of the sample being women. 27% of the sample identified themselves as being of non-white (European) ethnic origin. Of the total sample, 12.2% identified themselves as being of Indian origin, 6.1% of origin Black origin (i.e. Afro-Caribbean, Black African, etc.) 3.1% of Pakistani and 1% of Bangladeshi origin, with the remaining 5.6% of the sample being from a variety of different ethnic origins (e.g. Pacific Islander, Maori, etc.).

3.5 CRTB2: RELIABILITY

3.5.1 RELIABILITY: INTERNAL CONSISTENCY

Table 1 presents alpha coefficients for the CRTB2 subtests on a number of different samples. Inspection of this table indicates that all these coefficients are above .8, indicating that the CRTB2 has good levels of internal consistency reliability.

3.5.2 RELIABILITY: 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. Theoretically, test-retest and internal consistency estimates of a test's reliability should be equivalent to each other. The principal difference between the test-retest reliability statistic and the alpha coefficient is that the latter provides an estimate of the lower bound of the test's reliability and, as such, the alpha coefficient for any given test is usually lower than the test-retest reliability statistic.

Table 1 – Alpha coefficients for the Verbal (VCR2) and Numerical (NCR2) Critical Reasoning Subtests.

	Insurance Sales Agents (n=132)	MBA Students (n=205)	Undergraduates (n=70)	Professional/ Managerial (n=201)	Graduate/ Managerial (n=857)
VCR2	.88	.84	.88	.87	.82
NCR2	.83	.81	.86	.81	.88

Table 2 – Correlations between the Verbal (VCR2) and Numerical (NCR2) Critical Reasoning Subtests.

Insurance Sales Agents (n=132)	MBA Students (n=205)	Undergraduates (n=70)	Professional/ Managerial (n=201)	Graduate/ Managerial (n=857)
.40	.57	.49	.54	.52

3.6 CRTB2: CONSTRUCT VALIDITY

3.6.1 THE RELATIONSHIP BETWEEN THE VCR2 AND THE NCR2

Table 2 presents the correlations between the VCR2 and NCR2 on a variety of different samples. These data demonstrate that, as would be expected, while the Verbal and Numerical Critical Reasoning Tests are significantly correlated with each other, they are not so highly correlated with each other as to suggest that they are measuring the same construct. These correlations therefore provide support for both the convergent and discriminant construct validity of the VCR2 and the NCR2.

3.6.2 THE RELATIONSHIP BETWEEN THE CRBT2 AND THE APIL-B

The VCR2 and NCR2 were correlated with the APILB (Ability, Processing of Information and Learning Battery) that has been developed by Taylor (1995). The APIL-B has been specifically developed to be a culture fair assessment tool for use in a multi-racial context (South Africa). As such, it has been designed to assess an individual's core cognitive capabilities, rather than specific skills that may depend upon a person's educational experience and life advantages. Table 3 presents the correlations between the Verbal and Numerical Critical Reasoning tests and the APIL-B, on a sample of MBA students. These correlations are highly statistically significant, and substantial in size, providing strong support for the concurrent construct validity of the VCR2 and NCR2.

Table 3 – Correlations between the CRTB2 subtests and the APIL-B.

	APIL-B	Significance
VCR2	.57	n= 250, p<.001
NCR2	.51	n=169, p<.001

3.6.3 THE RELATIONSHIP BETWEEN THE CRBT2 AND THE GRT2

The Verbal and Numerical Critical Reasoning tests were, respectively, correlated with the verbal (r=.57, p<.001) and numerical (r=.51, p<.001) subtests of the General Reasoning Test (GRT2) on a sample of 25 undergraduates. These large, statistically significant correlations indicate that the VCR2 and NCR2 are measuring reasoning ability, rather than some other (related) construct such as verbal or numerical checking ability. These large correlations therefore provide strong support for the concurrent construct validity of the CRTB2.

3.6.4 THE RELATIONSHIP BETWEEN THE CRBT2 AND ABSTRACT REASONING ABILITY

To further assess whether the CRTB2 is measuring reasoning ability, rather than a related construct (such as checking ability) the VCR2 and NCR2 were correlated with the Abstract Reasoning subtest (AR2) of the GRT2, with the Ravens Advanced Progressive Matrices (RAPM) and with the Abstract Reasoning Test (ART). Each of these tests are measures of abstract reasoning ability (sometimes termed "fluid intelligence"). These data are presented in Tables 4, 5 and 6 respectively. All these correlations are statistically significantly and demonstrate that the critical reasoning tests are not simply measures of attention to numerical and verbal detail, but rather are measuring facets of general mental (reasoning) ability. These correlations therefore provide further support for the concurrent construct validity of the VCR2 and the NCR2.

Table 4 – Correlations between the CRTB2 subtests and the AR2.

	AR2	Significance
VCR2	.34	n= 985, p<.001
NCR2	.43	n= 985, p<.001

Table 5 – Correlations between the CRTB2 subtests and the RAPM.

	RAPM	Significance
VCR2	.29	n= 213, p<.001
NCR2	.48	n= 213, p<.001

Table 6 – Correlations between the CRTB2 subtests and the ART.

	ART	Significance

administration and handscoring.) The WAIS provides an overall IQ score, and two IQ subscale scores; termed Verbal and Performance IQ. The former assesses crystallised intelligence and the latter assess fluid intelligence. Correlations between the VCR2, the NCR2 and the MAB total score, subscale and subtest scores are presented in Table 7.

Both the VCR2 and the NCR2 were found to be significantly correlated with the MAB Verbal Subscale. As noted above, the Verbal (IQ) Subscale of the WAIS was designed to be a measure of crystallised intelligence rather than being a measure of verbal reasoning ability – as the subscale's name might (wrongly) suggest – and includes numerical (e.g. the Arithmetic subtest) as well as verbal subtests. As such, the substantial correlations between the MAB Verbal Subscale and the VCR2 and NCR2 provide strong support for the concurrent construct validity of both of the CRTB2 subtests, indicating that they are measuring aspects of crystallised intelligence (i.e. verbal/numerical reasoning ability).

Moreover, the significant correlation between the NCR2 and the MAB Performance subscale suggests that numerical critical reasoning ability weights more heavily on fluid intelligence than does verbal critical reasoning ability. (This is further suggested by the correlations reported above between the VCR2, the NCR2 and the various measures of abstract reasoning ability with which these tests have been correlated.) This finding is not surprising given the more abstract nature of the NCR2 items and thus provides further support for the concurrent construct validity of the NCR2.

Table 7 – Correlations between the CRTB2 subtests and the MAB.

	VCR2	NCR2
MAB Total	.24	.48
MAB Performance Subscale	.04	.48
MAB Verbal Subscale	.43	.44
MAB Verbal Subscale Subtests		
Information	.29	.32
Comprehension	.25	.44
Arithmetic	.24	.45
Similarities	.22	.33
Vocabulary	.32	.27
MAB Performance Subscale Subtests		
Digit Symbol	.09	.37
Picture Completion	.14	.38
Spatial	.19	.50
Picture Arrangement	.15	.34
Object Assembly	.09	.23

3.6.6 THE RELATIONSHIP BETWEEN THE CRBT2 AND INTELLECTANCE

Intellectance is a meta-cognitive variable that assesses a person's perception of their own level of mentality ability. While it is a personality factor, rather than an ability factor, Intellectance has nonetheless consistently been found to correlate with objective assessments of mental ability. As such it would be expected to be modesty correlated with critical reasoning ability. A sample (n=132) of insurance sales agents completed the VCR2 and NCR2 along with the 15FQ+. Intellectance (the 15FQ+ Factor ß) was found to correlate substantially (r=.42 and r=.36 respectively) with the VCR2 and the NCR2, thereby providing further support for the concurrent construct validity of the CRTB2.

3.6.7 THE RELATIONSHIP BETWEEN THE CRBT2 AND LEARNING STYLE

While learning style is not an ability factor, it has nonetheless been consistently found to be related to reasoning ability in meaningful ways. In particular, a more abstract rather than a more concrete learning style, and a more holistic rather than a more serial (i.e. focussing on the "big picture" rather than focussing on details) learning style, have been found to be associated with fluid intelligence but *not* with crystallised intelligence.

A sample of 134 respondents completed the VCR2, the NCR2 and the Learning Styles Inventory (LSI) for research purposes. The LSI scale Abstract-Concrete was found to be correlated with the NCR2 ($r=.26$, $p<.001$), but not with the VCR2. This reflects the fact, noted above, that numerical critical reasoning ability weights more heavily on fluid intelligence than does verbal critical reasoning ability. As such, this correlation provides further support for the concurrent construct validity of the CRTB2.

3.6.8 THE RELATIONSHIP BETWEEN THE CRBT2 AND EDUCATIONAL LEVEL

To further examine the concurrent construct validity of the CRTB2 the relationship between respondents' scores on this test and their educational level was examined. A one-way ANOVAs indicated that there were highly significant associations between the VRC2 ($F=10.3$, $df=4,686$ $p<.001$), the NCR2 ($F=29.2$, $df=4,686$ $p<.001$) and educational level. Table 8 presents mean scores on the CRTB2 subtests. (Means which share a common superscript are significantly different from each other at the 5% level – Tukey hsd). Inspection of this table indicates that respondents of higher educational level obtain on average higher scores on the VRC2 and NCR2 than do those of lower educational levels, thereby providing strong support for the concurrent construct validly of the CRTB

Table 8 – Mean VCR2 and NCR22 scores for respondents of each of the listed educational levels.

Completed Secondary Education		Trade/Technician Qualification		Tertiary Education		Degree		Post Graduate Degree	
VCR2	NCR2	VCR2	NCR2	VCR2	NCR2	VCR2	NCR2	VCR2	NCR2
18.6*	11.6*	19.3†	12.6†	22.7*†	14.9*†	22.0*†	16.3*†	23.2*†	16.7*†
(n=154)		(n=83)		(n=81)		(n=252)		(n=121)	

3.7 CRTB2: CRITERION VALIDITY

3.7.1 PREDICTING SALES SUCCESS

A sample of 132 Insurance Sales Agents completed the CRTB2 as part of a predictive criterion validation study. The association between their scores on the VCR2 and NCR2 and their job performance was examined using t-tests. Sales Agents were classified as either successful or unsuccessful depending upon their performance after one year in post. Table 9 presents the mean scores for these two groups on the VCR2 and NCR2. Inspection of this table indicates that, on average, the successful incumbents had higher scores on both of these tests than did the non-successful incumbents, with the difference in the scores of these two groups reaching statistical significance for the NCR2. This provides good support for the predictive criterion validity of this test, indicating that higher levels of critical reasoning ability are associated with better job performance.

Table 9 - Mean scores on the VCR2 and NCR2 for successful and unsuccessful insurance sales agents.

	VCR2	NCR2
Unsuccessful	18.1	12.6
Successful	21.2	18.2
	t=1.48 n.s.	t=2.18 p<.05

3.7.2 PREDICTING PRODUCTIVITY IN THE FINANCIAL SERVICES SECTOR

The performance of 98 insurance professionals was examined, with the staff being rated for their productivity over a 12 month period by their line-manager. The VCR2 and NCR2 were found to be significantly correlated with rated performance (r=.23, p<.05 and r=.27, p<.05 respectively). Not surprisingly, higher levels of both verbal and numerical critical reasoning ability were found to be associated with higher productivity. These data therefore provide strong support for the concurrent criterion validity of the CRTB2, demonstrating the potential utility of this measure for aiding selection and assessment decisions.

3.7.3 PREDICTING THE PERFORMANCE OF MBA STUDENTS

A group of MBA students completed the VCR2 and NCR2 prior to enrolling on their course. Their scores on these tests were then correlated with their subsequent performance across different courses on the MBA syllabus. The results of this analysis are presented in Table 10. Inspection of this table indicates that the critical reasoning tests were predictive of performance across a number of areas of study. These data provide strong support for the predictive criterion validity of the CRTB2.

Table 10 - Correlations between the VCR2, NCR2 and MBA performance.

	VCR2	NCR2
Innovation & design	.37† (n=89)	.26† (n=89)
Business decision making	.47† (n=35)	.43† (n=35)
Macro-economics	.48‡ (n=89)	.39‡ (n=89)
IT	.47† (n=35)	.51† (n=35)
Administration	.36* (n=34)	.51† (n=34)
Economics	.24 (n=56)	.01 (n=56)
Analytical Tools and Techniques	.31* (n=51)	.13 (n=51)
Marketing	.20 (n=53)	-.12 (n=53)
Finance & Accounting	.21 (n=56)	-.01 (n=56)
Organisational Behaviour	.30* (n=56)	-.03 (n=56)
Overall MBA Grade	.39† (n=48)	.11 (n=48)

*p<.05 †p<.01 ‡p<.001

3.7.4 PREDICTING MANAGERIAL PERFORMANCE

The Verbal and Numerical Critical Reasoning Tests were administered to 213 senior managers within the oil and gas industry as part of an assessment centre. Their performance on the assessment centre tasks was rated in a number of defined competencies, and their scores on the VCR2 and NCR2 were correlated with their rated performance on these competencies. These correlations are presented in Table 11, opposite. Inspection of this table indicates that critical reasoning ability was predictive of performance across a number of competencies, providing strong evidence of the concurrent criterion validity of the CRTB2. These data clearly demonstrate indicating that this instrument is a useful tool for aiding selection and assessment decisions.

Table 10 – Correlation between the CRTB2N, CRTB2V and each of the listed (rated) competencies.

	VCR2	NCR2
Communication	.21*	.33‡
Problem Solving	.44‡	.50‡
Leadership	.24†	.38‡
Decision Making	.28†	.31‡
Strategic Thinking	.28†	.34‡
Team Working	.18	.30†
Planning	.38‡	.33‡
Flexibility	.01	.23*
Commitment	.20	.32‡
Overall Competency Rating	.35‡	.48‡

*p<.05 †p<.01 ‡p<.001

3.8 CRTB2: 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 CRTB2 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.).

Ethnic bias in the CRTB2 items was examined on two international samples of 314 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 CRTB2 was examined on two international samples of 355 , 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 11 and 12 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 VCR2 and NCR2 respectively. Inspection of these tables indicates that the CRTB2 items show little bias by sex or ethnicity. 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 VCR2 items, a significance level of less than 0.001% would have had to have been adopted.

Table 11 – Item bias statistics for the VCR2.

Item No.	Mixed Ethnic vs. White European	Male vs. Female
1	ß=.142 p=.512	ß=.0.57 p=.726
2	ß=.031 p=.884	ß=.045 p=.525
3	ß=.072 p=.760	ß=.021 p=.451
4	ß=.566 p=.063	ß=.386 p=.076
5	ß=.175 p=.981	ß=.411 p=.019
6	ß=.383 p=.207	ß=.008 p=.871
7	ß=.119 p=.572	ß=.286 p=.073
8	ß=.077 p=.732	ß=.031 p=.477
9	ß=.109 p=.641	ß=.020 p=.764
10	ß=.212 p=.343	ß=.065 p=.671
11	ß=.249 p=.337	ß=.159 p=.237
12	ß=.129 p=.566	ß=.005 p=.006
13	ß=.007 p=.974	ß=.152 p=.361
14	ß=.003 p=.980	ß=.016 p=.679
15	ß=.039 p=.870	ß=.215 p=.209
16	ß=.029 p=.906	ß=.072 p=.658
17	ß=.014 p=.952	ß=.485 p=.036
18	ß=.141 p=.510	ß=.053 p=.522
19	ß=.216 p=.294	ß=.018 p=.648
20	ß=.322 p=.156	ß=.024 p=.611
21	ß=.206 p=.372	ß=.398 p=.022
22	ß=.204 p=.468	ß=.030 p=.529
23	ß=.501 p=.056	ß=.284 p=.102
24	ß=.191 p=.506	ß=.074 p=.714
25	ß=.279 p=.283	ß=.098 p=.322
26	ß=.012 p=.965	ß=.018 p=.595
27	ß=.071 p=.777	ß=.050 p=.590
28	ß=.092 p=.725	ß=.226 p=.214
29	ß=.432 p=.095	ß=.165 p=.044
30	ß=.112 p=.696	ß=.055 p=.514
31	ß=.068 p=.848	ß=.326 p=.044
32	ß=.072 p=.765	ß=.006 p=.863
33	ß=.127 p=.666	ß=.008 p=.988
34	ß=.183 p=.545	ß=.338 p=.067
35	ß=.313 p=.222	ß=.017 p=.773
36	ß=.039 p=.885	ß=.009 p=.901
37	ß=.677 p=.024	ß=.346 p=.067
38	ß=.540 p=.029	ß=.010 p=.760
39	ß=.066 p=.645	ß=.131 p=.444
40	ß=.317 p=.124	ß=.001 p=.981

Table 12 – Item bias statistics for the NCR2

Item No.	Mixed Ethnic vs. White European	Male vs. Female
1	ß=.174 p=.452	ß=.0.35 p=.615
2	ß=.117 p=.582	ß=.047 p=.284
3	ß=.277 p=.207	ß=.146 p=.382
4	ß=.181 p=.376	ß=.003 p=.944
5	ß=.174 p=.458	ß=.358 p=.194
6	ß=.385 p=.165	ß=.031 p=.720
7	ß=.016 p=.940	ß=.109 p=.518
8	ß=.574 p=.019	ß=.012 p=.778
9	ß=.435 p=.036	ß=.012 p=.726
10	ß=.470 p=.081	ß=.046 p=.262
11	ß=.150 p=.254	ß=.052 p=.721
12	ß=.278 p=.303	ß=.296 p=.131
13	ß=.177 p=.468	ß=.431 p=.025
14	ß=.538 p=.020	ß=.051 p=.396
15	ß=.194 p=.469	ß=.319 p=.114
16	ß=.175 p=.458	ß=.021 p=.819
17	ß=.385 p=.071	ß=.205 p=.210
18	ß=.154 p=.479	ß=.016 p=.806
19	ß=.384 p=.102	ß=.260 p=.134
20	ß=.242 p=.246	ß=.018 p=.736
21	ß=.263 p=.245	ß=.012 p=.752
22	ß=.006 p=.986	ß=011. p=.801
23	ß=.163 p=.496	ß=.009 p=.875
24	ß=.309 p=.186	ß=.020 p=.479
25	ß=.441 p=.061	ß=.018 p=.545

APPENDIX I – CRTB2 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; 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 CRTB2 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 ONLY ONE (either the Verbal or Numerical) OF THE CRITICAL REASONING TESTS IS BEING ADMINISTERED, then say:-

---

From now on, please do no 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 only one of the two tests contained in the booklet that I will shortly be distributing.

Say either:-

"The Verbal Critical Reasoning Test which takes 15 minutes."

---

or:-

"The Numerical Critical Reasoning Test which takes 25 minutes."

Continue by saying:-

"During the test I shall be checking to make sure 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."

IF YOU ARE ADMINISTERING BOTH THE VERBAL AND THE NUMERICAL CRITICAL REASONNG TESTS:-

Give an introduction as per or similar to that suggested above.

Then, continue by using the instructions exactly as given below. Say:-

---

"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. We shall be doing two tests, the Verbal Critical Reasoning Test which takes 15 minutes and the Numerical Critical Reasoning Test which takes 25 minutes. During the test I shall be checking to make sure 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:-

"Please note the answer boxes are in columns" (indicate) "*and remember do **not** write on the test booklet."*

"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."

IF YOU ARE ONLY DMINISTERING THE NUMERICAL CRITICAL REASONING TEST, then go to the section below headed Administering the Numerical Critical Reasoning Test.

IF YOU ARE ADMINISTERING BOTH CRITICAL REASONING TESTS, OR ARE JUST ADMINISTERING THE VERBAL CRITICAL REASONING TEST, 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, and continue:-

"In this test you have to draw inferences from short passages of text. You will be presented with a passage of text followed by a number of statements. Your task is to decide, on the basis of the information contained in the passage, whether each statement is true, false or cannot be inferred from the passage. Your decision should be based only on the information contained in the passage and not on your own knowledge or opinions."

"Mark your answer by filling in the appropriate box, on your answer sheet, that corresponds to your choice."

"You now have a chance to complete the example questions on page 3 in order to make sure that you understand the test. Enter your responses to the example questions in the section marked Example Questions at the top of the answer sheet."

Point to the section on the answer sheet marked Example Questions (as you read the above). Then pause while candidates read the instructions, then say:-

---

"Please attempt the example questions now."

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 no-one is looking at the actual test items, while completing the examples. When everyone has finished (allow a maximum of two and a half minutes) give the answers as follows:-

---

"The correct response to Example 1 is False. It is explicitly stated within the text that further growth in the number of radio stations is limited due to there being no new radio frequencies available."

"The correct response to Example 2 is True. It is explicitly stated that audience figures affect advertising revenue, thus affecting profitability."

"The correct response to Example 3 is Cannot Determine. It is impossible to infer, from the information provided in the text, whether radio stations in general will become more profitable. The text indicates that audience figures are currently poor for many radio stations and that it is expected that some may go bankrupt. However, it is not possible to infer from this that audience figures (and as a result advertising revenue) will increase for the remaining radio stations."

"Please do not turn over the page yet"

Check for understanding, then say:-

- *"Time is short, so when you begin the timed test work as quickly and as accurately as you can."*
- *"If you are unsure of an answer, mark your best choice and move on to the next question."*
- *"If you want to change an answer cross it out, as indicated in the instructions in the top left-hand corner of the answer sheet, and fill in your new choice of answer."*

Point to the top left-hand corner of the answer sheet as you read the above.

Then continue:-

- *"There are 8 passages of text and 40 questions. You have 15 minutes in which to answer them."*
- *"If you reach the end of the test 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 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 15 minutes say clearly:-

"Stop."

You should intervene if candidates continue after this point.

IF YOU ARE ONLY ADMINSITERING THE VERBAL CRITICAL REASONING TESTS, say:-

"Close the test booklets."

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

Then say:-

"Thank you for completing the Verbal Critical Reasoning Test."

IF YOU ARE ADMINISTERING BOTH OF THE CRITICAL REASONING TEST, continue by saying:-

" Now please turn to Page 12, which is a blank page."

Pause to allow the booklets to be turned to the correct page. 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?"

Rectify any omissions, 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, from the front of the group, and say:-

"Please turn to Page 14 of the booklet....."

IF YOU ARE ONLY ADMINISTERING THE NUMERICAL CRITICAL REASONING TEST, say:–

"Please open the booklet at Page 14...."

WHETHER YOU ARE ONLY ADMINISTERING THE NUMRICAL CRITIACL REASONING TEST, OR ARE ADMINSITERING BOTH TESTS, continue as follows:-

"...and follow the instructions for this test as I read them aloud."

"In this test you will have to draw inferences from numerical information which is presented in tabular form."

"You will be presented with a numerical table and asked a number of questions about this information. You will then have to select the correct answer to each question from one of six possible choices. One and only one answer is correct in each case."

"Mark your answer, by filling in the appropriate box, on your answer sheet that corresponds to your choice.

"You now have a chance to complete the example questions on Pages 15 in order to make sure that you understand the test. Enter your responses to the example questions in the section marked Example Questions at the top of the answer sheet."

Point to the section on the answer sheet marked Example Questions (as you read the above).

Pause while candidates read the instructions, then say:

"Please attempt the example questions now."

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 no-one is looking at the actual test items during the example session. When all have finished (allow a maximum of three minutes) give the answers as follows:

"The correct answer to Example 1 is Design (answer no. 5). It can be seen, in the table, that amongst women, design was consistently chosen by the lowest percentage as the most important feature of a car."

"The correct answer to Example 2 is performance (answer no. 1). It can be seen that of all the features of a car, performance is rated by men as being the most important feature of a car."

"The correct answer to Example 3 is 10.4 (answer no.5). Of men below the age of 30, 5% identified safety and 52% identified performance as the most important feature of a car. 52 over 5 is 10.4, therefore the answer is number 5."

"Please do not turn over the page yet"

Check for understanding, then say:-

- *"Time is short, so when you begin the timed test work as quickly and as accurately as you can."*
- *"If you are unsure of an answer, mark your best choice and move on to the next question."*
- *"If you want to change an answer cross it out, as indicated in the instructions in the top left-hand corner of the answer sheet, and fill in your new choice of answer."*

Point to the top left-hand corner of the answer sheet as you read the above.

Then continue:-

- *"There are 6 tables of numerical information and a total of 25 questions. You have 25 minutes in which to answer the questions."*
- *"If you reach the end of the test 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 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 25 minutes say clearly:-

"Stop. Close the test booklets"

You should intervene if candidates continue after this point.

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

……

IF YOU ARE ADMINISTERING BOTH OF THE CRITICAL REASONING TESTS, say:-

---

"Thank you for completing the Critical Reasoning Test Battery."

IF YOU ARE ONLY ADMINISTERING THE NUMERICAL CRITICAL REASONING TEST, say:-

---

"Thank you for completing the Numerical Critical Reasoning Test."

APPENDIX II – CRTB2 HAND SCORING INSTRUCTIONS

The completed answer sheets are scored and profiled by following the steps listed below:

1. Remove the top cover sheet of the combined answer/scoring sheet to reveal the scoring key.To score and standardise the VCR2 follow steps 2-8. To score and standardise the NCR2 follow steps 9-10.
2. Count up the number of correct responses for the VCR2 and enter the total (raw score) in the box marked "Total".

IF YOU DO NOT WISH TO CORRECT THE VCR2 SCORE FOR GUESSING GO STRAIGHT TO STEP 7.

3. To correct the VCR2 score for guessing add up the total number of incorrect responses (i.e. the total number of items attempted minus the raw score).
4. The correction for guessing can be found in Appendix III. The number of incorrect responses is listed in the first column of this table and the corresponding correction for guessing is listed in the second column. Make a note of the correction for guessing (that corresponds to the number of incorrectly completed items).
5. To obtain the corrected raw score, subtract the correction for guessing from the raw score. If this number is negative (i.e. the number corrected for guessing is larger than the raw score) then the corrected raw score is zero. Enter the corrected raw score in the box marked "Corrected/Uncorrected Raw Score". To indicate that you have made the correction, delete "Uncorrected".
6. To standardise the corrected raw score, look up the corrected raw score in Table 2 in Appendix IV and enter this in the box marked "Standard Score".

IF YOU DO NOT WISH TO CORRECT THE VCR2 SCORE FOR GUESSING CONTIUNE FROM STEP 7.

7. Enter the Total score obtained from step 2 in the box marked "Corrected/Uncorrected Raw score". To indicate that you have not made the correction, delete "Corrected".
8. To standardise the uncorrected raw score, look this value up in the Table 1 in Appendix IV and enter this in the box marked "Standard Score".

TO SCORE AND STANDARDISE THE NCR2 FOLLOW STEPS 9 - 10.

9. Count up the number of correct responses to the NCR2 and enter the total in the box marked "Total".
10. To standardise the raw score, look this value up in Table 3 in Appendix VI and enter this in the box marked "Standard Score".

APPENDIX III – CRTB2 SCORE CORRECTIONS FOR GUESSING

Table 1 – Correction for guessing

Number of incorrect answers	Correction - to be deducted from raw score
1	.5
2	1
3	1.5
4	2
5	2.5
6	3
7	3.5
8	4
9	4.5
10	5
11	5.5
12	6
13	6.5
14	7
15	7.5
16	8
17	8.5
18	9
19	9.5
20	10
21	10.5
22	11
23	11.5
24	12
25	12.5
26	13
27	Corrected raw score = 0
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	

APPENDIX IV – NORM TABLES

Table 1 – NCR2 norm table.

NCR2 Raw Score	1-5	6-7	8-10	11-12	13-16	17-18	19-21	22-23	24-25
Stanine	1	2	3	4	5	6	7	8	9

Table 2 – VCR2 norm table for raw (uncorrected) scores.

VCR2 Raw Score	1-9	10-12	13-16	17-19	20-22	23-25	26-28	29-31	32-40
Stanine	1	2	3	4	5	6	7	8	9

Table 3 – VCR2 norm table for corrected scores.

VCR2 Corrected Score	0	1-4	5-8	9-11	12-15	16-19	20-23	24-27	28-40
Stanine	1	2	3	4	5	6	7	8	9

REFERENCES

Binet, A. (1910). *Les idees 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,* 209-224.

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

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the education process.* Boston, Heath.

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

Glaser, E. (1941). *An experiment in the development of critical thinking.* New York, J. J. Little and Ives Company.

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

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,* 262-274.

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.

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.*