Technical Report
UK Clinical Aptitude Test (UKCAT) Consortium
Testing Interval: 7 July 2009 – 10 October 2009
Executive Summary

Prepared by:

Brad Wu, Ph.D.

1 North Dearborn
Chicago, IL 60602

Non-disclosure and Confidentiality Notice

TABLE OF CONTENTS

1.0 BACKGROUND

The UK Clinical Aptitude Test (UKCAT) was administered in 2009 beginning on 7 July 2009 and ending 10 October 2009. In this period, a total of 23,721 exams were administered. The exam consisted of four cognitive subtests: Verbal Reasoning (VR), Quantitative Reasoning (QR), Abstract Reasoning (AR) and Decision Analysis (DA). Three forms each were developed for VR, QR and AR. DA employed two forms. The forms were developed from the operational items used in the previous administrations (from 2006 to 2008) and also from new items that were trialled in 2008. A fifth component, referred to as the Behavioural Test, was first piloted in the 2007 administration and is intended to assess non-cognitive attributes of empathy, integrity and robustness that are associated with good doctors and dentists. The behavioural tests were administered for research purposes and were not intended for use as part of the operational test; however, some general results were provided to candidates in the form of narrative descriptors of their trait characteristics. Four different behavioural instruments were implemented: MEARS (Managing Emotions and Resilience Scales), ITQ100 (Interpersonal Traits Questionnaire) or NACE (Narcissism, Aloofness, Confidence and Empathy), IVQ49 (Interpersonal Values Questionnaire) or MOJAC (a measure of ethical orientation) and SAI2 (Self Appraisal Inventory). In addition, abridged versions of ITQ (labeled ITQ50) and IVQ (labeled IVQ33) were combined and utilised; this combined version is labeled ITQ/IVQ.

Each exam consisted of a total of 175 items (162 operational and 13 pretest) for the cognitive tests and 49 to 125 items for the behavioural tests. The exam was administered via computer in a 120-minute time period. Examinees were given 90 minutes to complete the cognitive tests with each of the four tests timed separately. Thirty minutes were allotted for the behavioural section. Results were provided to the candidates at the conclusion of testing and later to schools to which the candidates had applied.

Design of Exam

The UKCAT is an aptitude exam and is designed to measure innate cognitive abilities, personality and learning styles. It is not an exam that measures student achievement. It does not contain any curriculum or science content. The four cognitive subtests are described below.

Verbal Reasoning Subtest

The Verbal Reasoning (VR) subtest consists of 44 items. There are 40 operational (scored) and 4 pretest (unscored) items on each form. Candidates are allowed 21 minutes to answer the 44 items. In addition, candidates are allotted one minute to read general instructions for the subtest.

The 44 items in the VR subtest are grouped into 11 testlets. Each testlet has 4 items that relate to a single reading passage. Items from 10 testlets are scored; items from one testlet (designated as pretest) are not scored. Testlets are randomly ordered for presentation to candidates. The four items within each testlet are also randomly ordered during administration. Note that candidates see all four items related to a passage (i.e., within a testlet) before they are presented with another passage with its four items.

Quantitative Reasoning Subtest

The Quantitative Reasoning (QR) subtest consists of 40 items. There are 36 operational (scored) and 4 pretest (unscored) items. Candidates are allowed 21 minutes to answer the 40 items. In addition, candidates are allotted one minute to read general instructions for the subtest.

Nine scored testlets and one unscored testlet are presented to the candidates. Each testlet contains four items related to the stimulus in the testlet (i.e., a graph, a table). Testlets are randomly ordered for presentation to candidates. The four items within each testlet are also randomly ordered during administration. As is the case with the VR subtest, candidates are administered all four items within a testlet before they are presented with the next testlet and its four items.

Abstract Reasoning Subtest

The Abstract Reasoning (AR) subtest consists of 65 items. There are 60 operational (scored) and 5 pretest (unscored) items. Candidates are allowed 15 minutes to answer the 65 items. In addition, candidates are allotted one minute to read general instructions for the subtest.

Twelve scored testlets and one unscored testlet are presented to the candidates. Each testlet contains five items related to the stimulus in the set (i.e., two images or configurations of polygons and symbols). Testlets are randomly ordered for presentation to candidates. The five items within each set are also randomly ordered during administration. All items within a testlet are administered before the next testlet is presented.

Decision Analysis Subtest

The Decision Analysis (DA) subtest consists of 26 items. All items are scored. There are no pretest items in the DA subtest. Candidates are allowed 29 minutes to answer the 26 items. In addition, candidates are allotted one minute to read general instructions for the subtest.

One testlet is presented to the candidates. The testlet contains 26 items related to the stimulus in the set (i.e., a scenario that contains various pages of text and perhaps tables). The 26 items within the testlet are presented in a pre-specified order.

As mentioned above, there are no pretest items for the DA subtest. All items used in the DA subtest were pretested as a separate testing event (i.e., as a pretest study, not part of the live exam) in November and December of 2008.

2.0 EXAMINEE PERFORMANCE

Examinees' scale scores were reported for each cognitive subtest and were based on all the scored items for each section. The valid scale score ranged from 300 to 900, with a mean set to 600 in the 2006 reference sample. Universities received the subtest scaled scores for each candidate, plus a total score that is a simple sum of the four subtest scores and that had a valid range of 1200 to 3600.

An Item Response Theory (IRT) calibration model and IRT true score equating methods were used to transform the raw scores on each form onto a common reporting scale.

Table 1 presents summary statistics for each of the subtests, plus the total scale score for the 2009 UKCAT population. A total of 23,721 candidate scores were collected during the 2009 testing window and were used in these analyses. The scale score means varied across the four subtests. The mean scale score was 582.36 for VR, 637.77 for QR, 606.82 for AR and 677.62 for DA. Standard deviations ranged from 46.33 (DA) to 86.92 (QR). The distributions are generally symmetric around their means and reasonably well spread out.

Average scale score performance on VR for the total group was roughly equivalent to that of 2008. The mean QR and AR scale scores increased slightly. The mean score for DA showed a larger increase and, therefore, led to an increase in total scale score average compared to the previous year. The increase of DA average score, while large, was not alarming when taking into account the standard error of measurement. Because the error band for DA was quite large in 2008, less confidence can be placed on the 2008 DA scores and also comparison of 2009 to 2008 DA scores. The performance patterns for different subgroups (ethnic, gender, age and NS-SEC) closely paralleled that of the previous year. The majority of the group differences were not statistically significant.

Unlike the cognitive sections, no numeric result was provided to candidates after completion of the behavioural test. For each behavioural test, ordered categories were developed and scores for each test were classified into one of five categories. Cut-points on the scores used to make these classifications were obtained in two different ways. For the ITQ, IVQ and SAI2 tests, the score scales were cut at 5th, 30th, 70th and 95th percentiles based on the author's classification. For MEARS the score cuts were provided by Team Focus and represented the 10th, 30th, 70th, and 90th percentiles of a sample of data collected by Team Focus. Candidates were provided only the narrative description of the categories corresponding to their scores. Under the cut scores that were applied to assign narrative descriptors, nearly all candidates were clustered into the top two categories on the SAI2 tests, while classification of the ITQ, IVQ and MEARS scores showed spreads close to a normal distribution with small variation.

Table 2 shows the summary statistics for the Behavioural subtests. Total scores of the behavioural tests were all normally distributed with varying degrees of spread.
Aside from SAI2, distributions of behavioural categories also approximated normal. For SAI2, the distribution was slightly skewed, with most candidates falling into the high categories. The classification scheme of SAI2 should be regarded as exploratory as it was newly introduced to the UKCAT and not used previously on this population. Analyses of behavioural test scores by gender, ethnicity, NS-NEC, and age subgroups revealed insignificant differences between groups for all the tests.

3.0 TEST AND ITEM ANALYSIS

Test analysis for the operational forms included computation of the raw and scale score means, standard deviations, internal consistency reliabilities and standard error of measurement (SEM) of each form of each subtest. Item analysis included a complete classical analysis of item characteristics including *p* values, corrected point-biserial and biserial correlations (indices of item discrimination). IRT analyses included estimation of item parameters and standard errors. The IRT parameter estimates were re-scaled to be comparable with the previous years.

Test Analysis

Table 3 provides the raw score means, standard deviations, ranges, internal consistency reliabilities (Cronbach's alpha) and SEM for each form of each subtest. The means were similar across all forms within each subtest, with the exception of AR where the low and high means differed by approximately 9 points. The highest raw score reliabilities were found in AR, which can be attributed to the test length. Raw Score internal consistency reliabilities were .69, .68 and .66 for the three VR forms; .77, .76 and .67 for QR; .81, .83 and .82 for AR; and .59 and .65 for DA.

SEM were on the raw score metric and were approximately 2.9 for QR (number of items = 40), approximately 2.8 for QR (number of items = 36), 3.3-3.6 for AR (number of items = 60) and approximately 2.3 for DA (number of items = 26). The score reliability pattern in 2009 resembled that of 2008 and ranged from moderate to high.

Because scale scores (not raw scores) are the scores that are reported to candidates, scale score reliabilities and standard errors are also provided. Table 4a contains the scale score reliabilities and SEM for each form of the cognitive tests. Unlike the raw score reliability, where the reliability index (Cronbach's alpha) was generated based on the intercorrelations or internal consistency among the items, the overall reliability of the scale scores depends on the conditional reliability at each scale score point instead of on item scores. For this reason, the two reliability indices (Cronbach's alpha and marginal reliability of scale scores) are not directly comparable. The results indicate that scale score reliabilities ranged from moderate to high for VR, QR and AR. Scale score reliabilities ranged from .69-.74 for the VR forms and 73-.78 for the QR forms. Again, reliabilities for the AR subtests were higher (.85-.89) and better reflected the range of reliabilities desired for large-scale testing because of the length. The moderate reliability coefficients for the DA scale scores (.63-.68) was a result of the shorter test length (26 items). However, because of the small standard deviations, the two DA forms had small SEM (23-29) regardless of the moderate reliability coefficients. The SEM for DA in 2009 was considerably lower than 2008, where the SEM for DA ranged from 60-68. This indicates that the DA scores are more reliable in 2009 than those in 2008. SEM averaged around 43 for VR, 41 for QR and 28 for AR.

Table 4b contains the reliabilities and SEM for the total scale score. These values were computed as a composite function of the standard errors and reliabilities of the cognitive test forms contributing to the total. That is, each total scale score is a simple sum (linear composite) of the four forms of the cognitive tests that a given candidate was administered. There were 6 different combinations of cognitive test forms and, therefore, there were 6 different estimates of total scale score reliability and SEM. The range of values and the means are reported. The average reliability for total scale score was .89, reflecting high reliability. The average SEM was 96.89, which is quite reasonable for the range of total scale score.

In summary, score reliabilities of the four cognitive subtests in the 2009 UKCAT ranged from moderate to high. SEM for the subtest and total scores was satisfactory. Variation in score reliability and SEM across the four subtests can be attributed to test length, range of discrimination and difficulty among items.

Item Analysis

Item characteristics were examined based on Classical Test Theory and Item Response Theory. Both operational and pretest items were analysed.

For the cognitive sections, the results of the operational item analyses differed from the 2008 results in the overall quality of the pool. Range of difficulty and item discrimination were considerably better in 2009 across the VR, QR, AR and DA subtests. The pretest statistics, however, were very similar to those of 2008 and generally had poorer statistics. This is mostly because of the smaller sample for pretest items. However, pretest statistics usually improve as they are operationalised and reanalysed based on much larger samples. Item statistics from previous administrations were used not only for screening and item bank management. They were reviewed carefully and provided to item developers for the improvement of future item writing. Several item reviewing and writing workshops were arranged, and new pretest items were developed to comply with the improved guidelines. These items will be trialled in the 2010 administration and included in the new active item pool for future test construction.

Item-level results for the behavioural tests can be summarised as follows:

1. The IVQ tests (IVQ33 and IVQ49) and the MEARS subscales (Cognitive, Emotional and Behavioural) had very strong item-total correlations, indicating good discrimination power. The ITQ tests (ITQ100 and ITQ50) and SAI2 showed slightly lower item-total correlations, but all within acceptable range.

2. ITQ test items correlated consistently in the correct pattern (i.e., Narcissism and Aloofness items were negatively correlated with total score, but were positively correlated with Empathy and Confidence items). In terms of magnitude, only about 5-6% of the ITQ intercorrelations had absolute values smaller than .1. Generally speaking, the ITQ appears to be less internally consistent with respect to the total score. However, the ITQ is comprised of four subsections, and as such the total score is a multidimensional composite. Under these circumstances, the item total correlations would be expected to be lower than those from single construct measures.

Construct Validity

Internal construct validity refers to the degree to which the items in a test are related to the scale(s) that they are intended to measure and not related to the scale(s) that they are not explicitly intended to measure. Internal construct validity, evaluated through item-total correlations with scales and subscales, provided strong evidence that most items were measuring consistently within the expected scale structures. While this level of validity evidence does not address the criterion-related validity that is of primary interest for these tests, the findings reported here provide some foundational validity evidence for continued usage of these tests.

Table 7 contains the correlations among the behavioural and cognitive tests using scale scores. Most of the correlations between the behavioural and cognitive tests were small (absolute value < .10) and most were negative. The strongest relationships occurred between the ITQ and IVQ tests with Verbal Reasoning. In general, these values indicate very weak relationships between the behavioural and cognitive tests. The value of this cannot be determined at this point, but the finding that the behavioural tests did not appear to have a lot in common with the cognitive tests leaves open the possibility that they may contribute useful information in a predictive sense. Criterion-related analyses will be needed to evaluate whether the behavioural tests are related to performance in medical school or more generally to performance in practice. If they are, the possibility remains open that they may serve as a useful adjunct to the cognitive tests for predicting future performance.

4.0 DIFFERENTIAL ITEM FUNCTIONING

Introduction

Differential Item Functioning (DIF) refers to the potential for items to behave differently for different groups. DIF is generally an undesirable characteristic of an item because it means that the item is measuring both the construct it was designed to measure and some additional characteristic or characteristics of performance that depend on classification or membership in a group, usually a gender or ethnic group classification. For instance, if female and male examinees of the same ability level perform very differently on an item, then the item may be measuring

something other than the ability of the examinees, possibly some aspect of the examinees that is related to gender. The principles of test fairness require that examinations undergo scrutiny to detect and remove items that behave in significantly different ways for different groups based solely on these types of demographic characteristics. In DIF, the terms “reference group” and “focal group” are used for group comparisons and generally refer to the *majority* and the *minority* demographic groupings of the exam population.

This section describes the methods used to detect DIF for the UKCAT and provides the results for the 2009 administration.

Detection of DIF

There are a number of different procedures that can be used to detect DIF, and one of the most frequently used is the Mantel-Haenszel procedure. The Mantel-Haenszel procedure compares reference and focal group performance for examinees within the same ability strata. If there are overall differences between reference group and focal group performance for examinees of the same ability levels, then the item may not be fitting the psychometric model and may be measuring something other than what it was designed to measure.

The Mantel-Haenszel procedure requires a criterion of proficiency or ability that can be used to match (group) examinees into various levels of ability. For the UKCAT, matching is done using the raw score on each subtest associated with the item under study.

Items were classified for DIF using the Mantel-Haenszel delta statistic. This DIF statistic (hereafter known as MH D-DIF) is expressed as *differences* on the delta scale, which is commonly used to indicate the difficulty of test items. For example, a MH D-DIF value of 1.00 means that one of the two groups being analysed found the question to be one delta point more difficult than did *comparable* members of the other group. (Except for extremely difficult or easy items, a difference of one delta point is approximately equal to a difference of 10 points in percent correct between groups). We have adopted the convention of having negative values of MH D-DIF reflect an item that is differentially more difficult for the focal group (generally, females or the ethnic minority group). Positive values of MH D-DIF indicate that the item is differentially more difficult for the reference group (generally white or male candidates). Both positive and negative values of the DIF statistic are found and are taken into account by these procedures.

Criteria for Flagging Items

For the UKCAT, MH DIF items will be classified into one of three categories, A, B, or C. Category A contains items with negligible DIF, Category B contains items with slight to moderate DIF, and Category C contains items with moderate to large DIF. These categories are derived from the DIF classification categories developed by Educational Testing Service (ETS) and are defined below:

A: MH D-DIF is not significantly different from zero or has an absolute value < 1.0
B: MH D-DIF is significantly different from zero and has an absolute value >= 1.0 and < 1.5
C: MH-D-DIF is significantly larger than 1.0 and has an absolute value >= 1.5.

The scale units are based on a delta transformation of the proportion correct measure of item difficulty. The delta for an item is defined as: delta = 4z + 13, where z is the z-score that cuts off p (the proportion correct for an item) in the standard normal distribution. The delta scale removes some of the non-linearity of the proportion correct scale and allows easier interpretation of classical item difficulties.

Items flagged in Category C are typically subjected to further scrutiny. Items flagged in Category A are not reviewed, while Category B items may be reviewed. The principal interpretation of

Category C items is that items flagged in this category, based on the present samples, appear to be functioning differently for the reference and focal groups under comparison. If an item functions differently for two different groups, then content experts may (or may not) be able to determine from the item itself whether the item text contains language or content that may create a bias for the reference or focal group. Therefore, Category C flagging for DIF is necessary but not sufficient grounds for revision and possible removal of the item from the pools.

Comparison Groups for DIF Analysis

DIF analyses were conducted for the pretest and operational items when sample sizes were large enough. The UKCAT DIF comparison groups are based on gender, age, ethnicity and SEC.

Male is treated as the reference group and female as the focal group.

Age was separated into groups less than 20 years old and greater than 35 years old. The age group less than 20 was considered the reference group and the group greater than 35 was considered the focal group.

There are 17 ethnic categories in the UKCAT database. For the DIF analyses, several of these categories were collapsed into meaningful larger groups. The "White" group was treated as the reference group and all other minority groups were focal groups. The DIF ethnic categories used for these analyses (collapsed where indicated) were as follows:

White. White – British, White – Irish, White – Other.
Black. Black – Black/British – African, Black – Black/British – Caribbean, Black – Black/British Other.
Asian. Chinese, Asian – Asian/British – Bangladeshi, Asian – Asian/British – Indian, Asian – Asian/British – Other Asian, Asian – Asian/British – Pakistani.
Mixed. Mixed – Mixed – Other, Mixed – White/Asian, Mixed – White/Black African, Mixed – White/Black Caribbean.
Other. Other ethnic group.
Information Withheld.

For DIF analysis on SEC, comparisons were examined only between SEC Class 1 and other Classes (Class 2 to 5) because of the limited sample sizes in Classes 2 to 5. SEC Class 1 was the majority and therefore considered the reference group. All other SEC Classes were treated as focal groups.

Sample Size Requirements

Minimum sample size requirements used for the UKCAT DIF analyses were at least 50 focal group candidate responses and at least 400 total (focal plus reference) candidate responses. Because pretest items are distributed across multiple versions of the forms, fewer responses are available per item than for operational items. As a result, it was not possible to compute DIF for many of the pretest items for some group comparisons.

DIF Results

Tables 5 and 6 show the number and percentages of items classified into each of the three DIF categories along with the numbers for which insufficient data was available to compute DIF (Category NA). The results for the operational items are given in Table 5. Those for the pretest items are in Table 6.

In operational DIF analysis, all items met sample size requirements to compute DIF for all subtests and comparison groups. For pretest items, only the comparisons between age groups, between white and those who withheld information and between SEC Classes 1 and 4 did not meet minimal sample size requirements. The percent of items not meeting sample size requirements ranged from 3% to 7% for age comparison. Only one item in a subtest did not meet the sample requirement for the White/Withheld Information and SEC Class 1/4 comparisons. These items failed to meet the minimal sample requirement due to the relatively small samples collected in the focal groups (age > 35, ethnic information withheld and SEC Class 4). These items will be re-evaluated for DIF when they are used in future operational forms.

For the operational pools (Table 5), there were 23 occurrences of Category C DIF across all cognitive subtests and comparisons. The average proportion of Category C DIF out of all possible comparisons across the four cognitive tests was 0.45%. Of these 23 occurrences, 11 occurred for the Age <20/>35 comparison, 4 for the White/Black comparison, 5 for the White/Other comparison, 2 for the White/Withheld Information comparison and 1 for the SEC Class 1/5 comparison. Items with Category C DIF will be reviewed in the future item writing workshops, where content and wording will be examined for the potential of bias against specific groups. These items will either be revised or retired based on the review in the item writing workshops.

For the pretest items, there were 10 occurrences of Category C DIF: 2 for the Male/Female comparison, 2 for the White/Black comparison, 3 for the White/Mixed comparison, 2 for the White/Other comparison and 1 for the SEC Class 1/3 comparison. The proportion of Category C DIF out of all possible comparisons across the three cognitive pretests was 0.39%. Pretest items showing Category C DIF will also be reviewed in the item writing workshops and the decision of revision or retirement will be made based on the content of the items. Taken together, the results indicate very little DIF occurrence in the UKCAT items.

Of the 33 overall occurrences (operational and pretest) of Category C DIF, 21 favored the reference group; of the remaining 12 items favoring the focal groups, most (8 items) were from the Age <20/>35 comparison, one was from the White/Black comparison, one was from the White/Mixed comparison and two were from the White/Other comparison.

While items showing Category C DIF were reviewed in previous item writing workshops, results of DIF analysis had not been used for item screening from 2006 to 2008 because of the limited number of active items in the pool. As the number of active items increase through years of pretesting, DIF statistics can be considered as an additional item screening criterion. Items continuously showing significant DIF may need to be retired from the pool to ensure the overall quality of the item pool.

5.0 REFERENCES

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 207-210.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple group IRT analysis and test maintenance for binary items [Computer program]*. Chicago: Scientific Software International.

6.0 TABLES

Table 1: Subtest and Total Scale Score Summary Statistics: Total Population

Test	Total N	Mean	Standard Deviation	Minimum	Maximum
Verbal Reasoning	23721	582.36	79.17	300	900
Quantitative Reasoning	23721	637.77	86.92	300	900
Abstract Reasoning	23721	606.82	79.65	300	900
Decision Analysis	23721	677.62	46.33	300	900
Total Scale Score	23721	2504.57	220.35	1200	3340

Table 2: Behavioural Subtest and Total Scale Score Summary Statistics: Total Population

Test	Total N	Mean	Standard Deviation	Minimum	Maximum
ITQ-100	4790	289.28	20.23	212	359
IVQ-49	4772	118.79	14.39	74	170
ITQ50	4738	142.36	10.71	95	184
IVQ-33	4737	79.87	9.98	23	111
MEARS Cognitive	4701	183.83	20.51	76	243
MEARS Behavioural	4701	185.32	21.50	86	245
MEARS Emotional	4701	111.25	11.97	45	143
SAI2	4720	233.74	19.65	108	284

Table 3: Raw Score Test Statistics

Test	Form	N Items	N Candidates	Mean	SD	Min	Max	Alpha	SEM
Verbal Reasoning	1	40	3461	25.77	5.41	6	39	0.69	2.99
	2	40	10054	24.30	5.26	0	40	0.68	2.96
	3	40	10206	25.70	5.12	3	39	0.66	2.99
Quantitative Reasoning	1	36	3461	19.88	5.83	3	36	0.77	2.80
	2	36	10054	18.42	5.78	0	36	0.76	2.84
	3	36	10206	15.81	5.04	1	34	0.67	2.89
Abstract Reasoning	1	60	3461	33.03	8.30	4	58	0.81	3.62
	2	60	10054	42.30	8.02	0	60	0.83	3.31
	3	60	10206	37.14	8.44	3	58	0.82	3.61
Decision Analysis	1	26	12196	12.78	3.63	1	24	0.59	2.33
	2	26	11525	16.11	3.81	0	26	0.65	2.25

Table 4: Scale Score Reliability and Standard Error of Measurement for Cognitive Subtests

Tests	Form	N Items	N Candidates	Mean	SD	Min	Max	Scale Score Reliability	SEM
Verbal Reasoning	1	40	3461	594.21	80.94	310	890	0.74	41.59
	2	40	10054	580.96	79.80	300	900	0.70	43.71
	3	40	10206	579.73	77.58	300	890	0.69	43.12
Quantitative Reasoning	1	36	3461	679.85	84.17	380	900	0.78	39.48
	2	36	10054	641.09	88.07	300	900	0.78	41.40
	3	36	10206	620.22	81.22	320	900	0.73	42.36
Abstract Reasoning	1	60	3461	604.85	78.51	300	900	0.86	29.27
	2	60	10054	611.26	79.13	300	900	0.85	30.24
	3	60	10206	603.11	80.34	300	890	0.89	26.28
Decision Analysis	1	26	12196	663.84	46.67	350	810	0.63	28.58
	2	26	11525	692.21	41.22	300	900	0.68	23.42

Table 4b: Scale Score Reliability and Standard Error of Measurement for Total Scale Score

Reliability		SEM	
Range*	Mean	Range	Mean
.86 - .91	.89	92.38 – 101.7	96.89

* Based on 6 combinations of cognitive test forms.

Table 5: DIF Classification. Operational Pool

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Analysis	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent	Count	Percent
Male/Female	A	118	98.33%	107	99.07%	178	98.89%	52	100.00%
	B	2	1.67%	1	0.93%	2	1.11%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	NA*	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
Age <20/>35	A	112	93.33%	101	93.52%	166	92.22%	40	76.92%
	B	7	5.83%	4	3.70%	11	6.11%	8	15.38%
	C	1	0.83%	3	2.78%	3	1.67%	4	7.69%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
White/Black	A	109	90.83%	95	87.96%	176	97.78%	48	92.31%
	B	10	8.33%	10	9.26%	4	2.22%	4	7.69%
	C	1	0.83%	3	2.78%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
White/Asian	A	119	99.17%	106	98.15%	180	100.00%	52	100.00%
	B	1	0.83%	2	1.85%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
White/mixed	A	119	99.17%	106	98.15%	178	98.89%	51	98.08%
	B	1	0.83%	2	1.85%	2	1.11%	1	1.92%
	C	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
White/other	A	113	94.17%	100	92.59%	177	98.33%	49	94.23%
	B	7	5.83%	5	4.63%	2	1.11%	2	3.85%
	C	0	0.00%	3	2.78%	1	0.56%	1	1.92%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
White/Wthld. Inf.	A	117	97.50%	107	99.07%	173	96.11%	51	98.08%
	B	2	1.67%	0	0.00%	7	3.89%	1	1.92%

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Analysis	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent	Count	Percent
	C	1	0.83%	1	0.93%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
SEC Class 1/2	A	118	98.33%	107	99.07%	178	98.89%	52	100.00%
	B	2	1.67%	1	0.93%	2	1.11%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
SEC Class 1/3	A	118	98.33%	106	98.15%	180	100.00%	52	100.00%
	B	2	1.67%	2	1.85%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
SEC Class 1/4	A	118	98.33%	108	100.00%	178	98.89%	51	98.08%
	B	2	1.67%	0	0.00%	2	1.11%	1	1.92%
	C	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%
SEC Class 1/5	A	120	100.00%	107	99.07%	176	97.78%	51	98.08%
	B	0	0.00%	1	0.93%	3	1.67%	1	1.92%
	C	0	0.00%	0	0.00%	1	0.56%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	108	100.00%	180	100.00%	52	100.00%

*NA: Insufficient data to compute MH D-DIF

Table 6: DIF Classification. Pretest Pool

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent
Male/Female	A	71	98.61%	70	97.22%	85	94.44%
	B	1	1.39%	1	1.39%	4	4.44%
	C	0	0.00%	1	1.39%	1	1.11%
	NA*	0	0.00%	0	0.00%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%
Age <20/>35	A	69	95.83%	67	93.06%	87	96.67%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	3	4.17%	5	6.94%	3	3.33%
	Total	72	100.00%	72	100.00%	90	100.00%
White/Black	A	71	98.61%	71	98.61%	86	95.56%
	B	1	1.39%	1	1.39%	2	2.22%
	C	0	0.00%	0	0.00%	2	2.22%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%
White/Asian	A	66	91.67%	69	95.83%	88	97.78%
	B	6	8.33%	3	4.17%	2	2.22%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%
White/mixed	A	71	98.61%	70	97.22%	90	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	1	1.39%	2	2.78%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%
White/other	A	71	98.61%	69	95.83%	86	95.56%
	B	1	1.39%	1	1.39%	4	4.44%
	C	0	0.00%	2	2.78%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%
White/Wthld. Inf.	A	71	98.61%	71	98.61%	89	98.89%
	B	0	0.00%	0	0.00%	0	0.00%

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent
	C	0	0.00%	0	0.00%	0	0.00%
	NA	1	1.39%	1	1.39%	1	1.11%
	Total	72	100.00%	72	100.00%	90	100.00%
SEC Class 1/2	A	70	97.22%	71	98.61%	89	98.89%
	B	2	2.78%	1	1.39%	1	1.11%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%
SEC Class 1/3	A	70	97.22%	69	95.83%	86	95.56%
	B	2	2.78%	2	2.78%	4	4.44%
	C	0	0.00%	1	1.39%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%
SEC Class 1/4	A	72	100.00%	71	98.61%	90	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	1	1.39%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%
SEC Class 1/5	A	72	100.00%	72	100.00%	90	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	72	100.00%	72	100.00%	90	100.00%

*NA: Insufficient data to compute MH D-DIF

Table 7: Correlations of Cognitive Scale Scores and Behavioural Tests

		Verbal Reasoning	Quantitative Reasoning	Abstract Reasoning	Decision Analysis	ITQ100	IVQ49	ITQ55	IVQ33	MEARS Cognitive	MEARS Behavioural	MEARS Emotional	SAI2
Verbal	Pearson Correlation	1.000	0.505	0.330	0.423	0.092	-0.082	0.058	-0.078	0.021	-0.039	0.008	-0.014
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.148	0.008	0.561	0.351
	N	23721	23721	23721	23721	4790	4771	4738	4737	4697	4697	4697	4720
Quantitative	Pearson Correlation	0.505	1.000	0.414	0.432	-0.003	-0.098	0.008	-0.076	0.015	-0.035	-0.028	-0.023
	Sig. (2-tailed)	0.000		0.000	0.000	0.835	0.000	0.581	0.000	0.304	0.016	0.058	0.118
	N	23721	23721	23721	23721	4790	4771	4738	4737	4697	4697	4697	4720
Abstract	Pearson Correlation	0.330	0.414	1.000	0.381	0.006	-0.070	0.023	-0.061	0.010	0.006	0.004	0.022
	Sig. (2-tailed)	0.000	0.000		0.000	0.693	0.000	0.109	0.000	0.490	0.688	0.783	0.132
	N	23721	23721	23721	23721	4790	4771	4738	4737	4697	4697	4697	4720
Decision	Pearson Correlation	0.423	0.432	0.381	1.000	0.034	-0.066	0.050	-0.083	0.019	-0.009	-0.005	0.000
	Sig. (2-tailed)	0.000	0.000	0.000		0.017	0.000	0.001	0.000	0.201	0.535	0.749	0.992
	N	23721	23721	23721	23721	4790	4771	4738	4737	4697	4697	4697	4720
ITQ100	Pearson Correlation	0.092	-0.003	0.006	0.034	1.000	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	0.000	0.835	0.693	0.017		.	.	.	.	.	.	.
	N	4790	4790	4790	4790	4790	0	0	0	0	0	0	0
IVQ49	Pearson Correlation	-0.082	-0.098	-0.070	-0.066	.(a)	1.000	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	.		.	.	.	.	.	.
	N	4771	4771	4771	4771	0	4771	0	0	0	0	0	0
ITQ55	Pearson Correlation	0.058	0.008	0.023	0.050	.(a)	.(a)	1.000	0.309	.(a)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	0.000	0.581	0.109	0.001	.	.		0.000	.	.	.	.
	N	4738	4738	4738	4738	0	0	4738	4737	0	0	0	0
IVQ33	Pearson Correlation	-0.078	-0.076	-0.061	-0.083	.(a)	.(a)	0.309	1.000	.(a)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	.	.	0.000		.	.	.	.
	N	4737	4737	4737	4737	0	0	4737	4737	0	0	0	0
Cognitive	Pearson Correlation	0.021	0.015	0.010	0.019	.(a)	.(a)	.(a)	.(a)	1.000	0.454	0.557	.(a)
	Sig. (2-tailed)	0.148	0.304	0.490	0.201	.	.	.	.		0.000	0.000	.
	N	4697	4697	4697	4697	0	0	0	0	4697	4697	4697	0
Behavioural	Pearson Correlation	-0.039	-0.035	0.006	-0.009	.(a)	.(a)	.(a)	.(a)	0.454	1.000	0.390	.(a)
	Sig. (2-tailed)	0.008	0.016	0.688	0.535	.	.	.	.	0.000		0.000	.
	N	4697	4697	4697	4697	0	0	0	0	4697	4697	4697	0
Emotional	Pearson Correlation	0.008	-0.028	0.004	-0.005	.(a)	.(a)	.(a)	.(a)	0.557	0.390	1.000	.(a)
	Sig. (2-tailed)	0.561	0.058	0.783	0.749	.	.	.	.	0.000	0.000		.
	N	4697	4697	4697	4697	0	0	0	0	4697	4697	4697	0
SAI2	Pearson Correlation	-0.014	-0.023	0.022	0.000	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	1.000
	Sig. (2-tailed)	0.351	0.118	0.132	0.992	.	.	.	.	.	.	.	
	N	4720	4720	4720	4720	0	0	0	0	0	0	0	4720

(a) Cannot be computed because at least one of the variables is constant