



Technical Report
UK Clinical Aptitude Test (UKCAT) Consortium
Testing Interval: 6 July 2010 – 8 October 2010
Executive Summary

Prepared by:
Brad Wu, Ph.D.

1 North Dearborn
Chicago, IL 60602



Non-disclosure and Confidentiality Notice

This document contains confidential information concerning Pearson's services, products, data security procedures, data storage parameters, and data retrieval processes. You are permitted to view and retain this document provided that you disclose no part of the information contained herein to any outside agent or employee, except those agents and employees directly charged with reviewing this information. These employees should be instructed and agree not to disclose this information for any purposes beyond the terms stipulated in the agreement of your company or agency with Pearson.

Copyright © 2010 Pearson, Inc. All rights reserved. PEARSON logo is a trademark in the U.S. and/or other countries.

TABLE OF CONTENTS

1.0	BACKGROUND	4
	Design of Exam	4
	<i>Verbal Reasoning Subtest</i>	4
	<i>Quantitative Reasoning Subtest</i>	4
	<i>Abstract Reasoning Subtest</i>	5
	<i>Decision Analysis Subtest</i>	5
2.0	EXAMINEE PERFORMANCE	5
3.0	TEST AND ITEM ANALYSIS	6
	<i>Item Analysis</i>	7
	<i>Construct Validity</i>	8
4.0	DIFFERENTIAL ITEM FUNCTIONING	9
	<i>Introduction</i>	9
	<i>Criteria for Flagging Items</i>	9
	<i>Comparison Groups for DIF Analysis</i>	10
	<i>Sample Size Requirements</i>	10
	<i>DIF Results</i>	11
5.0	REFERENCES	12
6.0	TABLES	13
	Table 1: Subtest and Total Scale Score Summary Statistics: Total Population	13
	Table 2: Behavioural Subtest and Total Scale Score Summary Statistics: Total Population	13
	Table 3: Raw Score Test Statistics	14
	Table 4: Scale Score Reliability and Standard Error of Measurement for Cognitive Subtests	14
	Table 4b: Scale Score Reliability and Standard Error of Measurement for Total Scale Score	14
	Table 5: DIF Classification. Operational Pool	15
	Table 6: DIF Classification. Pretest Pool	17
	Table 7: Correlations of Cognitive Scale Scores and Behavioural Tests	19

1.0 BACKGROUND

The UK Clinical Aptitude Test (UKCAT) was administered in 2010 beginning on 6 July and ending 8 October. In this period, a total of 25,257 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 2009) and also from new items that were trialled in 2009. 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 instruments were used in 2010 to form 3 behavioural tests : MEARS (Managing Emotions and Resilience Scales), a combined ITQ50 (Interpersonal Traits Questionnaire) and IVQ33 (Interpersonal Values Questionnaire), and SAI2 (Self Appraisal Inventory).

Each exam consisted of a total of 173 items (158 operational and 15 pretest) for the cognitive tests and 83 to 125 items for the behavioural tests. The exam was administered via computer in a 120-minute time period. Examinees were given 93 minutes to complete the cognitive tests with each of the four tests timed separately. Twenty-seven 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 36 items. There are 32 operational (scored) and 4 pretest (unscored) items. Candidates are allowed 22 minutes to answer the 40 items. In addition, candidates are allotted one minute to read general instructions for the subtest.

Eight 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 28 items. Twenty-six of the 28 items are scored. Two new items were created and pretested in 2010 to provide some room for recycling and replacement. Candidates are allowed 31 minutes to answer the 28 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 28 items related to the stimulus in the set (i.e., a scenario that contains various pages of text and perhaps tables). The 28 items within the testlet are presented in a pre-specified order.

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 2010 UKCAT population. While scale score means varied across the four subtests, distributions are generally symmetric around their means and reasonably well spread out. The mean scale score for VR and AR stay fairly close to the previous years (2006-2009). For QR and DA, fluctuation in average scale scores was observed. This is due to the structural changes implemented in those two tests. Whenever major structural changes are applied to a test, previous parameters (e.g., item difficulty) may no longer fit the new structure and therefore resulting in score shift. Once new parameters are obtained based on the modified test condition, they will then be used for rescaling, so future scores will regress to the reference scale.

The score shift and regression described above can be observed in DA section from the past three years. The average scale score for DA was 618.53 in 2008. When brand new forms were introduced in 2009, the average DA scale score was shifted up to 677.62 because a new benchmark was established. Using the 2009 data, the item parameters were re-estimated based on a much larger sample and used to scale the 2010 DA forms. In 2010, the average score was reverted back to 615.55.

For QR, structural changes such as abridged scored section and additional timing were implemented in 2010. These changes set a new benchmark for QR and therefore shifted the average score from 637.77 in 2009 up to 673.30 in 2010. The item parameters were re-estimated using the 2010 data. These new parameters will then be used to scale 2011 QR forms. As in the case of DA, a score regression back to the lower 600s will be expected in 2011.

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 ITQ50/IVQ33 and SAI2 tests, the score scales were cut at 5th, 30th, 70th and 95th percentiles based on the test developer's (TUNRA) 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 skewed, with most candidates falling into the highest two categories. The classification scheme of SAI2 should be regarded as exploratory as fundamental difference might exist between the UKCAT population and the population used to establish the classification scheme, which possibly caused the skewed distribution observed.

Analyses of behavioural test scores by gender, ethnicity, NS-SEC, 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 DA where raw score means of the two forms differed by approximately 3 points. The highest raw score reliabilities were found in AR, which can be attributed to the test length. SEM were on the raw score metric and were approximately 3.0 for QR (number of items = 40), approximately 2.6 for QR (number of items = 32), 3.4 for AR (number of items = 60) and approximately 2.3 for DA (number of items = 26). The score reliability pattern in 2010 showed slight improvement compared to previous years (2006-2009) 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 inter-correlations 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 comparable. The results indicate that scale score reliabilities ranged from moderate to high for all cognitive tests. As in the raw score reliability, scale score reliabilities for the AR forms were higher (.85-.87) 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 (.66 and .68) were a result of the shorter test length (26 items).

In addition to test length, quality of scored items (e.g., item discrimination power) can also affect overall score reliability. The effect was observed in the QR section, where the test was shortened yet score reliabilities increased from 2009 to 2010. The improvement can be partially attributed to the higher average discrimination among the scored items selected in 2010.

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 .88, reflecting high reliability. The average SEM was 97.83, which is quite reasonable for the range of total scale score.

In summary, score reliabilities of the four cognitive subtests in the 2010 UKCAT ranged from moderate to high. Reliability for the total score was satisfactory. Variation in score reliability across the four tests can be partially attributed to the length of subtests. Improvement of score reliability compared to previous years, however, is a result of a stronger item bank and thus higher flexibility in selecting better fitted (more discriminative and reasonably challenging) 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 2009 results in the overall quality of the pool. Range of difficulty and item discrimination were considerably better in 2010 across the VR, QR, AR and DA subtests. The pretest statistics, however, were very similar to those of 2009 and generally had poorer statistics. This is mostly because of the smaller sample collected 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, but also 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 2011 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. IVQ33 and the MEARS subscales (Cognitive, Emotional and Behavioural) had very strong item-total correlations, indicating good discrimination power. ITQ50 and SAI2 showed slightly lower item-total correlations, but all within acceptable range.
2. ITQ50 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 correlations between ITQ50 items and the subscales had absolute values smaller than .1. Generally speaking, ITQ50 appeared to be less internally consistent with respect to the total score. However, ITQ50 is comprised of four subscales, 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 ITQ50 and Verbal Reasoning. In general, these values indicate very weak relationships between the behavioural and cognitive tests. 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 2010 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: $\text{delta} = 4z + 13$, 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, some comparisons between age groups, between white and mixed race, other race, those who withheld information, and SEC classes did not meet minimal sample size requirements. Only about 1% of all possible comparisons did not meet the sample size requirements, with the highest frequency observed in the age group comparisons. These comparisons failed to meet the minimal sample requirements due to the relatively small samples collected in the focal groups (e.g., age > 35 and ethnic information withheld). These items will be re-evaluated for DIF when they are used in future operational forms.

For the operational pools (Table 5), there were 16 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 less than 0.4%. Of these 16 occurrences, 9 occurred in the Age <20/>35 comparison, 2 in the White/Black comparison, and 5 in the White/Other comparison. No other group comparisons showed signs of significant DIF. For the pretest items, there were 33 occurrences of Category C DIF, which was less than .8% of all comparisons. While the number of Category C DIF identified in the 2010 pretest pool was slightly larger than 2009, the size of the pretest pool was also larger in 2010. Thus, the proportion of Category C DIF did not change significantly from 2009 to 2010. Taken together, the results indicate very little DIF occurrence in the UKCAT items.

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). *BLOG-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

{tc "Report " \f C \ 1}{tc "Detailed and/or summarized report " \f C \ 2}Test	Total N	Mean	Standard Deviation	Minimum	Maximum
Verbal Reasoning	25257	574.32	78.70	300	890
Quantitative Reasoning	25257	673.30	95.69	300	900
Abstract Reasoning	25257	625.80	91.27	300	900
Decision Analysis	25257	615.55	102.77	300	900
Total Scale Score	25257	2488.96	285.08	1330	3520

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

{tc "Report " \f C \ 1}{tc "Detailed and/or summarized report " \f C \ 2}Test	Total N	Valid Range	Mean	Standard Deviation	Minimum	Maximum
ITQ50	8338	48-192	142.34	10.75	95	181
IVQ33	8338	30-120	80.19	9.89	36	119
MEARS Cognitive	8400	41-246	184.55	20.46	87	240
MEARS Behavioural	8400	42-252	186.15	21.71	91	245
MEARS Emotional	8400	24-144	111.86	12.12	47	144
SAI2	8406	72-288	234.03	19.93	137	285



Table 3: Raw Score Test Statistics

Test	Form	N Items	N Candidates	Mean	SD	Min	Max	Alpha	SEM
Verbal Reasoning	1	40	8840	23.59	5.62	4	39	0.70	3.08
	2	40	8117	23.79	5.35	2	38	0.68	3.03
	3	40	8300	23.21	5.35	1	38	0.68	3.04
Quantitative Reasoning	1	32	8840	16.53	5.36	0	32	0.75	2.66
	2	32	8117	17.28	5.52	0	32	0.79	2.51
	3	32	8300	16.91	5.18	1	32	0.75	2.57
Abstract Reasoning	1	60	8840	39.90	8.28	0	60	0.82	3.46
	2	60	8117	39.09	8.59	3	60	0.83	3.51
	3	60	8300	41.24	8.82	6	60	0.85	3.37
Decision Analysis	1	26	12864	13.70	3.82	0	25	0.63	2.33
	2	26	12393	16.77	3.79	1	26	0.65	2.23

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	8840	575.22	79.77	300	890	0.73	41.76
	2	40	8117	579.79	79.82	300	890	0.72	42.39
	3	40	8300	568.00	75.96	300	890	0.69	42.09
Quantitative Reasoning	1	32	8840	668.97	97.42	300	900	0.77	46.72
	2	32	8117	679.92	95.10	300	900	0.79	43.37
	3	32	8300	671.42	94.07	330	900	0.77	45.50
Abstract Reasoning	1	60	8840	625.15	90.15	300	900	0.85	34.68
	2	60	8117	614.48	89.91	300	900	0.85	34.35
	3	60	8300	637.55	92.35	300	900	0.87	33.93
Decision Analysis	1	26	12864	617.58	108.00	300	900	0.66	62.70
	2	26	12393	613.45	97.01	300	900	0.68	55.30

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

Reliability		SEM	
Range*	Mean	Range	Mean
.85 - .91	.88	93.57 – 106.34	97.83



* Based on 6 combinations of cognitive test forms.

Table 5: DIF Classification. Operational Pool

Comparison Group	MH-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Analysis	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent
Male/Female	A	88	100.00%	87	98.86%	155	100.00%	52	100.00%
	B	0	0.00%	1	1.14%	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	88	100.00%	88	100.00%	155	100.00%	52	100.00%
Age <20/>35	A	77	87.50%	80	90.91%	147	94.84%	41	78.85%
	B	7	7.95%	7	7.95%	5	3.23%	10	19.23%
	C	4	4.55%	1	1.14%	3	1.94%	1	1.92%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	88	100.00%	88	100.00%	155	100.00%	52	100.00%
White/Black	A	86	97.73%	83	94.32%	151	97.42%	51	98.08%
	B	2	2.27%	4	4.55%	4	2.58%	0	0.00%
	C	0	0.00%	1	1.14%	0	0.00%	1	1.92%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	88	100.00%	88	100.00%	155	100.00%	52	100.00%
White/Asian	A	88	100.00%	85	96.59%	155	100.00%	52	100.00%
	B	0	0.00%	3	3.41%	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	88	100.00%	88	100.00%	155	100.00%	52	100.00%
White/mixed	A	87	98.86%	86	97.73%	150	96.77%	52	100.00%
	B	1	1.14%	2	2.27%	5	3.23%	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	88	100.00%	88	100.00%	155	100.00%	52	100.00%
White/other	A	84	95.45%	75	85.23%	150	96.77%	50	96.15%



Comparison Group	MH-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Analysis	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent
	B	4	4.55%	9	10.23%	4	2.58%	2	3.85%
	C	0	0.00%	4	4.55%	1	0.65%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	88	100.00%	88	100.00%	155	100.00%	52	100.00%
White/Wthld. Inf.	A	85	96.59%	85	96.59%	150	96.77%	47	90.38%
	B	3	3.41%	3	3.41%	5	3.23%	5	9.62%
	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	88	100.00%	88	100.00%	155	100.00%	52	100.00%
SEC Class 1/2	A	88	100.00%	88	100.00%	153	98.71%	52	100.00%
	B	0	0.00%	0	0.00%	2	1.29%	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	88	100.00%	88	100.00%	155	100.00%	52	100.00%
SEC Class 1/3	A	88	100.00%	88	100.00%	155	100.00%	52	100.00%
	B	0	0.00%	0	0.00%	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	88	100.00%	88	100.00%	155	100.00%	52	100.00%
SEC Class 1/4	A	88	100.00%	88	100.00%	155	100.00%	52	100.00%
	B	0	0.00%	0	0.00%	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	88	100.00%	88	100.00%	155	100.00%	52	100.00%
SEC Class 1/5	A	88	100.00%	86	97.73%	154	99.35%	51	98.08%
	B	0	0.00%	2	2.27%	1	0.65%	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	88	100.00%	88	100.00%	155	100.00%	52	100.00%

*NA: Insufficient data to compute MH D-DIF



Table 6: DIF Classification. Pretest Pool

Comparison Group	MH-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Analysis	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent
Male/Female	A	120	97.56%	103	90.35%	143	95.33%	4	100.00%
	B	3	2.44%	8	7.02%	6	4.00%	0	0.00%
	C	0	0.00%	3	2.63%	1	0.67%	0	0.00%
	NA*	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
Age <20/>35	A	115	93.50%	103	90.35%	147	98.00%	4	100.00%
	B	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	1	0.67%	0	0.00%
	NA	8	6.50%	11	9.65%	2	1.33%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
White/Black	A	122	99.19%	110	96.49%	148	98.67%	3	75.00%
	B	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	C	1	0.81%	4	3.51%	2	1.33%	1	25.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
White/Asian	A	107	86.99%	100	87.72%	139	92.67%	3	75.00%
	B	13	10.57%	12	10.53%	9	6.00%	1	25.00%
	C	3	2.44%	2	1.75%	2	1.33%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
White/mixed	A	119	96.75%	112	98.25%	149	99.33%	4	100.00%
	B	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	C	1	0.81%	1	0.88%	1	0.67%	0	0.00%
	NA	3	2.44%	1	0.88%	0	0.00%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
White/other	A	120	97.56%	113	99.12%	148	98.67%	3	75.00%
	B	0	0.00%	0	0.00%	0	0.00%	1	25.00%
	C	1	0.81%	0	0.00%	0	0.00%	0	0.00%



Comparison Group	MH-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Analysis	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent
	NA	2	1.63%	1	0.88%	2	1.33%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
White/Wthld. Inf.	A	119	96.75%	111	97.37%	149	99.33%	3	75.00%
	B	0	0.00%	0	0.00%	0	0.00%	1	25.00%
	C	0	0.00%	1	0.88%	0	0.00%	0	0.00%
	NA	4	3.25%	2	1.75%	1	0.67%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
SEC Class 1/2	A	122	99.19%	113	99.12%	149	99.33%	4	100.00%
	B	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	C	1	0.81%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	1	0.88%	1	0.67%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
SEC Class 1/3	A	119	96.75%	114	100.00%	147	98.00%	4	100.00%
	B	3	2.44%	0	0.00%	2	1.33%	0	0.00%
	C	1	0.81%	0	0.00%	1	0.67%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
SEC Class 1/4	A	120	97.56%	113	99.12%	150	100.00%	4	100.00%
	B	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	C	3	2.44%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	1	0.88%	0	0.00%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	100.00%
SEC Class 1/5	A	121	98.37%	114	100.00%	149	99.33%	4	100.00%
	B	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	C	1	0.81%	0	0.00%	1	0.67%	0	0.00%
	NA	1	0.81%	0	0.00%	0	0.00%	0	0.00%
	Total	123	100.00%	114	100.00%	150	100.00%	4	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	ITQ55	IVQ33	MEARS Cognitive	MEARS Behavioural	MEARS Emotional	SAI2
Verbal	Pearson Correlation	1	0.554	0.355	0.490	0.091	-0.059	-0.023	-0.053	-0.046	-0.038
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.000	0.032	0.000	0.000	0.000
	N	25257	25257	25257	25257	8381	8380	8463	8463	8463	8412
Quantitative	Pearson Correlation	0.554	1	0.434	0.501	0.007	-0.099	-0.019	-0.038	-0.055	-0.028
	Sig. (2-tailed)	0.000		0.000	0.000	0.508	0.000	0.083	0.000	0.000	0.009
	N	25257	25257	25257	25257	8381	8380	8463	8463	8463	8412
Abstract	Pearson Correlation	0.355	0.434	1	0.439	0.026	-0.062	0.001	-0.033	0.003	0.018
	Sig. (2-tailed)	0.000	0.000		0.000	0.018	0.000	0.907	0.002	0.773	0.099
	N	25257	25257	25257	25257	8381	8380	8463	8463	8463	8412
Decision	Pearson Correlation	0.490	0.501	0.439	1	0.053	-0.063	-0.033	-0.047	-0.022	0.001
	Sig. (2-tailed)	0.000	0.000	0.000		0.000	0.000	0.002	0.000	0.041	0.916
	N	25257	25257	25257	25257	8381	8380	8463	8463	8463	8412
ITQ50	Pearson Correlation	0.091	0.007	0.026	0.053	1	0.307	.(a)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	0.000	0.508	0.018	0.000		0.000
	N	8381	8381	8381	8381	8381	8380	0	0	0	0
IVQ33	Pearson Correlation	-0.059	-0.099	-0.062	-0.063	0.307	1	.(a)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	
	N	8380	8380	8380	8380	8380	8380	0	0	0	0
Cognitive	Pearson Correlation	-0.023	-0.019	0.001	-0.033	.(a)	.(a)	1	0.461	0.571	.(a)
	Sig. (2-tailed)	0.032	0.083	0.907	0.002	.	.		0.000	0.000	.
	N	8463	8463	8463	8463	0	0	8463	8463	8463	0
Behavioural	Pearson Correlation	-0.053	-0.038	-0.033	-0.047	.(a)	.(a)	0.461	1	0.393	.(a)
	Sig. (2-tailed)	0.000	0.000	0.002	0.000	.	.	0.000		0.000	.
	N	8463	8463	8463	8463	0	0	8463	8463	8463	0
Emotional	Pearson Correlation	-0.046	-0.055	0.003	-0.022	.(a)	.(a)	0.571	0.393	1	.(a)
	Sig. (2-tailed)	0.000	0.000	0.773	0.041	.	.	0.000	0.000		.
	N	8463	8463	8463	8463	0	0	8463	8463	8463	0
SAI2	Pearson Correlation	-0.038	-0.028	0.018	0.001	.(a)	.(a)	.(a)	.(a)	.(a)	1
	Sig. (2-tailed)	0.000	0.009	0.099	0.916	
	N	8412	8412	8412	8412	0	0	0	0	0	8412

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