Technical Report
UK Clinical Aptitude Test Consortium
UKCAT Examination

Testing Interval: 7 July 2008 – 12 October 2008

Executive Summary

Prepared by:

Brad Wu, Ph.D.

1 North Dearborn
Chicago, IL 60602

DOCUMENT HISTORY

Version	Date	Description
1.0	6/22/2009	Document created by Brad Wu
2.0	6/30/2009	Document edited by Belinda Brunner

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.

TABLE OF CONTENTS

1.0 BACKGROUND

The UKCAT examination was administered in 2008 beginning on 7 July 2008 and ending 12 October 2008. In this period, a total of 20,511 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 items used in the 2007 administrations (obtained from Team Focus) and also from new items that had been previously trialled in 2006 and 2007. 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 test was administered for research purposes and was 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. Three different behavioural instruments were piloted: MEARS (Managing Emotions and Resilience Scales), ITQ100 (Interpersonal Traits Questionnaire) or NACE (Narcissism, Aloofness, Confidence and Empathy), and IVQ49 (Interpersonal Values Questionnaire) or MOJAC (a measure of ethical orientation). In addition, abridged versions of ITQ (labeled ITQ50) and IVQ (labeled IVQ33) were combined and piloted; this combined version is labeled ITQ/IVQ. One of the four behavioural subtests mentioned above was randomly assigned to an examinee along with the cognitive tests.

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 the Exam

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

Verbal Reasoning Subtest

The VR subtest consists of 44 total 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. Prior to taking the UKCAT, candidates are provided access to detailed instructions for all subtests and examples on the UKCAT website.

There are 11 testlets in the VR subtest. 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. Each testlet is randomly ordered for presentation to a candidate. The four items within each set 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 QR subtest consists of 40 total items. There are 36 operational (scored) and 4 pretest (unscored) items. Like the VR subtest, candidates are allowed 21 minutes to answer the 40 items. Similarly, 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). Each testlet is randomly ordered for presentation to a candidate. 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 AR subtest consists of 65 total items. There are 60 operational (scored) and 5 pretest (unscored) items. Candidates are allowed 15 minutes to answer the 65 items. Similar to the previous two subtests, candidates are allotted one minute to read general instructions for this 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). Each testlet is randomly ordered for presentation to a candidate. 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 DA subtest consists of 26 total items. All items are scored. There are no pretest items in the DA subtest. Candidates are allowed 29 minutes to answer the 26 items. As with the other subtests, candidates are allotted one minute to read general instructions for this 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. New items will be pretested as a separate testing event (i.e., as a pretest study, not part of the live exam).

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 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 cognitive subtests, plus the total summed scale score for the total group. A total of 20,511 exam scores collected through the 2008 testing window were used in these analyses. The mean scale score was 585.25 for VR, 629.51 for QR, 596.41 for AR, and 618.53 for DA. Standard deviations ranged from 83.66 (AR) to 102.91 (DA). The score distributions were 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 2007. The mean QR scale score dropped slightly. Mean scores for AR and DA increased somewhat and, therefore, the total scale score mean also increased. Performance for different subgroups (ethnic, gender, age and SC-NEC) closely paralleled that of the previous year.

Table 2 shows the summary statistics for the Behavioural subtests. A characteristic of all Behavioural subtests was that the score distributions were slightly concentrated and peaked. This characteristic was more obvious for ITQ and MEARS, where a portion of the lower score ranges were not present. In addition to the small numbers of candidates with low scores, a slight negative skew was also observed for MEARS.

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 and IVQ tests, the scale scores were cut at 5%, 30%, 70%, and 95% percentiles based on the test developer'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. Note that because these measures are still experimental, these cut scores should be regarded as preliminary. 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 classified into the same one or two categories on the MEARS tests, while classification of the ITQ/IVQ scores showed spreads close to a normal distribution with small variation. Analyses of behavioural test scores by gender, ethnicity, NS-NEC, and age subgroups revealed only slight differences between most groups for all the tests.

3.0 TEST AND ITEM ANALYSIS

Test analysis for the operational forms included computation of the scale score means, standard deviations, internal consistency reliabilities, and standard errors of measurement of each form of each subtest. Item analysis included a complete classical analysis of item characteristics including p-values, and 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 to the previous years.

Test Analysis

Table 3 provides the scale score means, standard deviations, ranges, internal consistency reliabilities (Cronbach's alpha), and standard errors of measurement for each form of each subtest. Cronbach's alpha is an internal consistency index that ranges from 0 to 1. The higher the index, the more reliable the test scores are. The value can also be affected by the length of the test. Thus interpretation of the value should also take into account the test length.

The results indicated that scale score reliabilities were a moderate .64-.66 for the VR forms, and .60-.63 for the QR. Reliabilities for the AR subtests were higher (.75-.81) and better reflect the range of reliabilities desired for large-scale testing. The lower reliabilities for the DA scale scores (.55-.61) were most likely the result of shorter test length (26 items) for that subtest. Standard errors were about 53 for VR, 49 for QR and 38 AR, and ranged from 60-68 for DA. These standard errors provide some guidance with respect to the importance placed on score differences (e.g., differences less than 1 standard error should not be regarded as meaningfully different).

Table 4 contains the reliabilities and standard errors 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 sections of the cognitive tests that a given candidate was administered. One of the multiple forms in each section was randomly assigned to each examinee. There were 6 different combinations of cognitive test forms and, therefore, there were 6 different estimates of total scale score reliability and standard error. The range of values and the means are reported in Table 4. The average reliability for total scale score was .86, reflecting good overall reliability. The average standard error was 103.47.

Score reliabilities of the behavioural subtests are presented in Table 5. The score reliabilities (Cronbach's Alpha) ranged from .741 for ITQ/IVQ to .944 for MEARS.

In sum, score reliabilities of the five sections in 2008 UKCAT ranged from moderate to high. Score reliability for the cognitive and the behavioural section were mostly satisfactory. Variation in score reliability across the four cognitive tests can be partially attributed to the length of subtests.

Item Analysis

Item characteristics were examined based on Classical Test Theory and Item Response Theory.

For the cognitive sections, the results of the item analyses differed from the 2007 results in terms of difficulty and discrimination power. Mean p-values, an index of item difficulty, were lower in 2008 across all subtests, except Abstract Reasoning pretest and Decision Analysis. Both of which remain fairly close to the mean scores of 2007. Mean point-biserials, an index of item discrimination power, were generally lower in 2008 (both operational and pretest) across Verbal Reasoning, Quantitative Reasoning, and Abstract Reasoning. As the available pool of newly pretested items increases, efforts will be made to increase the levels of point-biserials in new forms.

Item-level results for the behavioural tests can be summarized as follows: 1) the IVQ tests (IVQ33 and IVQ49), and the MEARS subscales (Cognitive, Emotional and Behavioural) had very strong item-total correlations, which indicated good discrimination power. 2) ITQ test items correlated consistently in the expected pattern (i.e. Narcissism and Aloofness items were negatively correlated with total score, while Empathy and Confidence correlated positively with total score).
3) Generally speaking, ITQ and MEARS appeared to be less internally consistent in their measurement with respect to the total score. However, it must be recognized that these tests are comprised of several 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, such as IVQ.

Construct Validity

Table 6 contains the correlations among the behavioural and cognitive test scores. Higher correlations can be observed between the cognitive sections. Correlations between the behavioural and cognitive tests were generally low (absolute value < .10) and most were negative. This finding suggests that cognitive and behavioural sections are tapping quite different constructs. Therefore, the behavioural tests may contribute useful additional information in a predictive sense. Criterion-related analyses will be needed to evaluate whether the behavioural tests are related to performance in medical and dental school or more generally to performance in practice.

Internal construct validity, evaluated through correlations between item scores and scale/subscale scores, provided strong evidence that the behavioural test 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 provide supporting evidence for continued research using behavioural tests.

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 examination because it means that the test 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. The terms “reference group” and “focal group” are used in DIF for group comparisons and generally refer to the “majority” and the “minority” demographic groupings for the exam population.

This section describes the methods used to detect DIF for the UKCAT and provides the results for the 2008 administration.

Detection of DIF

There are a number of different procedures that can be used to detect differential item functioning, 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 the 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 analyzed 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 ETS DIF classification categories 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 for DIF.

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, and ethnicity. Age was separated into groups < 20 years old and > 35 years old. There are 17 ethnic categories in the UKCAT database. For the DIF analyses several of these categories were collapsed into meaningful larger 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

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 + 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 7 and 8 show the number and percentages of items classified into each of the three DIF categories along with the numbers for which insufficient data were available to compute DIF (category NA). The results for the operational items are given in Table 7, those for the pretest items in Table 8.

In operational DIF analysis, age comparison within the Quantitative Reasoning had seven items that failed to meet the sample size requirement because fewer than 50 cases were found in the age group >35. All other operational items met sample size requirements to compute DIF for all subtests and comparison groups. For the pretest items, the percent of items not meeting sample size requirements ranged from 0% for male/female and White/Asian to 100% for Age <20/>35 and White/Withheld information. The substantial number of non-qualifying items was due the relatively small samples collected on the pretest items. These items will be re-evaluated for DIF as they make their way into future operational forms.

For the operational pools (Table 7) 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 0.6%. Of these 16 occurrences, 8 occurred for Age <20/>35 comparison, 3 for the White/Black comparison, 3 for White/Other comparison, 1 for the White/Mixed comparison, and 1 for White/Withheld Information comparison. No other comparison groups showed signs of important DIF. For the pretest items there were 7 occurrences of category C DIF, 3 for White/Black, 2 for White/Asian, and 2 for White/Other. Taken together, the results indicate very little DIF occurring 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). *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

Test	Total N	Mean	Standard Deviation	Minimum	Maximum
Verbal Reasoning	20511	585.25	88.62	300	880
Quantitative Reasoning	20511	629.51	96.78	300	900
Abstract Reasoning	20511	596.41	83.66	300	900
Decision Analysis	20511	618.53	102.91	300	900
Total Scale Score	20511	2429.70	274.63	1210	3380

Table 2: Score Summary of the Behavioural Tests

Test	Total N	Valid Range	Mean	Standard Deviation	Minimum	Maximum
ITQ100	4557	96-384	289.18	18.96	201	360
IVQ49	4397	45-180	117.28	14.32	39	172
ITQ50	4661	48-192	142.06	10.03	106	183
IVQ33	4661	30-120	79.18	10.03	30	116
MEARS Cognitive	6896	41-246	184.41	20.08	41	241
MEARS Behavioural	6896	42-252	184.84	21.00	42	249
MEARS Emotional	6896	24-144	110.06	11.63	24	142

Table 3: 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	7041	589.93	92.36	300	880	0.65	52.56
	2	40	6604	581.13	83.78	300	880	0.66	51.13
	3	40	6866	584.41	89.04	300	880	0.64	54.06
Quantitative Reasoning	1	36	7041	636.67	107.64	300	900	0.60	51.13
	2	36	6604	640.14	90.76	300	900	0.61	49.33
	3	36	6866	611.95	87.82	300	900	0.63	47.41
Abstract Reasoning	1	60	7041	605.63	81.10	300	900	0.81	35.03
	2	60	6604	581.92	81.30	300	890	0.75	41.84
	3	60	6866	600.88	86.59	300	900	0.79	37.34
Decision Analysis	1	26	10279	617.54	99.91	300	900	0.61	59.58
	2	26	10232	619.53	105.84	300	880	0.55	67.56

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

Reliability		SEM	
Range*	Mean	Range	Mean
.84 - .87	.86	100.57 – 106.60	103.47

* Based on 6 combinations of cognitive test forms

Table 5: Score Reliability Indices for the Behavioural Subtests

Item Statistics	N Items	N Candidates	Minimum	Maximum	Reliability Coefficient (Cronbach's Alpha)
ITQ100	96	4557	201	360	.783
IVQ49	45	4397	39	172	.903
ITQ50/IVQ33	78	4661	136	299	.741
MEARS	125	6896	107	632	.944

Table 6: Correlations of Cognitive Scale Scores and Behavioural Test Scores

		Verbal Reasoning	Quantitative Reasoning	Abstract Reasoning	Decision Analysis	ITQ100	IVQ49	ITQ55	IVQ33	MEARS Cognitive	MEARS Behavioural	MEARS Emotional
Verbal Reasoning	Pearson Correlation	1	.452(**)	.323(**)	.418(**)	.056(**)	-.076(**)	.072(**)	-.090(**)	-.003	-.072(**)	-.032(**)
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.831	.000	.008
	N	20511	20511	20511	20511	4557	4397	4661	4661	6896	6896	6896
Quantitative Reasoning	Pearson Correlation	.452(**)	1	.361(**)	.394(**)	-.029	-.053(**)	-.017	-.093(**)	.036(**)	-.021	-.016
	Sig. (2-tailed)	.000		.000	.000	.051	.000	.249	.000	.003	.079	.177
	N	20511	20511	20511	20511	4557	4397	4661	4661	6896	6896	6896
Abstract Reasoning	Pearson Correlation	.323(**)	.361(**)	1	.391(**)	-.011	-.007	.011	-.075(**)	.027(*)	-.009	.019
	Sig. (2-tailed)	.000	.000		.000	.447	.665	.459	.000	.028	.442	.122
	N	20511	20511	20511	20511	4557	4397	4661	4661	6896	6896	6896
Decision Analysis	Pearson Correlation	.418(**)	.394(**)	.391(**)	1	-.001	-.061(**)	.022	-.135(**)	.005	-.047(**)	-.022
	Sig. (2-tailed)	.000	.000	.000		.937	.000	.136	.000	.688	.000	.073
	N	20511	20511	20511	20511	4557	4397	4661	4661	6896	6896	6896
ITQ100	Pearson Correlation	.056(**)	-.029	-.011	-.001	1	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	.000	.051	.447	.937		.	.	.	.	.	.
	N	4557	4557	4557	4557	4557	0	0	0	0	0	0
IVQ49	Pearson Correlation	-.076(**)	-.053(**)	-.007	-.061(**)	.(a)	1	.(a)	.(a)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	.000	.000	.665	.000	.		.	.	.	.	.
	N	4397	4397	4397	4397	0	4397	0	0	0	0	0
ITQ55	Pearson Correlation	.072(**)	-.017	.011	.022	.(a)	.(a)	1	.240(**)	.(a)	.(a)	.(a)
	Sig. (2-tailed)	.000	.249	.459	.136	.	.		.000	.	.	.
	N	4661	4661	4661	4661	0	0	4661	4661	0	0	0
IVQ33	Pearson Correlation	-.090(**)	-.093(**)	-.075(**)	-.135(**)	.(a)	.(a)	.240(**)	1	.(a)	.(a)	.(a)
	Sig. (2-tailed)	.000	.000	.000	.000	.	.	.000		.	.	.
	N	4661	4661	4661	4661	0	0	4661	4661	0	0	0
MEARS Cognitive	Pearson Correlation	-.003	.036(**)	.027(*)	.005	.(a)	.(a)	.(a)	.(a)	1	.431(**)	.540(**)
	Sig. (2-tailed)	.831	.003	.028	.688	.	.	.	.		.000	.000
	N	6896	6896	6896	6896	0	0	0	0	6896	6896	6896
MEARS Behavioural	Pearson Correlation	-.072(**)	-.021	-.009	-.047(**)	.(a)	.(a)	.(a)	.(a)	.431(**)	1	.333(**)
	Sig. (2-tailed)	.000	.079	.442	.000	.	.	.	.	.000		.000
	N	6896	6896	6896	6896	0	0	0	0	6896	6896	6896
MEARS Emotional	Pearson Correlation	-.032(**)	-.016	.019	-.022	.(a)	.(a)	.(a)	.(a)	.540(**)	.333(**)	1
	Sig. (2-tailed)	.008	.177	.122	.073	.	.	.	.	.000	.000	
	N	6896	6896	6896	6896	0	0	0	0	6896	6896	6896

* Correlation is significant at the 0.05 level (2-tailed)
** Correlation is significant at the 0.01 level (2-tailed)
(a) Cannot be computed because at least one of the variables is constant

Table 7: 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	117	97.5	106	98.1	179	99.4	51	98.1
	B	3	2.5	2	1.9	1	0.6	1	1.9
	C	0	0.0	0	0.0	0	0.0	0	0.0
	NA*	0	0.0	0	0.0	0	0.0	0	0.0
	Total	120	100.0	108	100.0	180	100.0	52	100.0
Age <20/>35	A	111	92.5	95	88.0	167	92.8	47	90.4
	B	5	4.2	5	4.6	10	5.6	5	9.6
	C	4	3.3	1	0.9	3	1.7	0	0.0
	NA	0	0.0	7	6.5	0	0.0	0	0.0
	Total	120	100.0	108	100.0	180	100.0	52	100.0
White/Black	A	111	92.5	94	87.0	178	98.9	43	82.7
	B	8	6.7	13	12.0	2	1.1	8	15.4
	C	1	0.8	1	0.9	0	0.0	1	1.9
	NA	0	0.0	0	0.0	0	0.0	0	0.0
	Total	120	100.0	108	100.0	180	100.0	52	100.0
White/Asian	A	118	98.3	106	98.1	180	100.0	49	94.2
	B	2	1.7	2	1.9	0	0.0	3	5.8
	C	0	0.0	0	0.0	0	0.0	0	0.0
	NA	0	0.0	0	0.0	0	0.0	0	0.0
	Total	120	100.0	108	100.0	180	100.0	52	100.0
White/mixed	A	118	98.3	106	98.1	177	98.3	52	100.0
	B	1	0.8	2	1.9	3	1.7	0	0.0
	C	1	0.8	0	0.0	0	0.0	0	0.0
	NA	0	0.0	0	0.0	0	0.0	0	0.0
	Total	120	100.0	108	100.0	180	100.0	52	100.0
White/other	A	113	94.2	94	87.0	177	98.3	45	86.5
	B	6	5.0	12	11.1	3	1.7	7	13.5
	C	1	0.8	2	1.9	0	0.0	0	0.0
	NA	0	0.0	0	0.0	0	0.0	0	0.0
	Total	120	100.0	108	100.0	180	100.0	52	100.0
White/Wthld. Inf.	A	114	95.0	105	97.2	175	97.2	50	96.2
	B	5	4.2	3	2.8	5	2.8	2	3.8
	C	1	0.8	0	0.0	0	0.0	0	0.0

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Analysis	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent	Count	Percent
	NA	0	0.0	0	0.0	0	0.0	0	0.0
	Total	120	100.0	108	100.0	180	100.0	52	100.0

*NA: Insufficient data to compute MH D-DIF

Table 8: DIF Classification. Pretest Pool

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent
Male/Female	A	70	97.2	69	95.8	88	97.8
	B	2	2.8	3	4.2	2	2.2
	C	0	0.0	0	0.0	0	0.0
	NA*	0	0.0	0	0.0	0	0.0
	Total	72	100.0	72	100.0	90	100.0
Age <20/>35	A	0	0.0	0	0.0	0	0.0
	B	0	0.0	0	0.0	0	0.0
	C	0	0.0	0	0.0	0	0.0
	NA	72	100.0	72	100.0	90	100.0
	Total	72	100.0	72	100.0	90	100.0
White/Black	A	56	77.8	29	40.3	72	80.0
	B	1	1.4	0	0.0	1	1.1
	C	3	4.2	0	0.0	0	0.0
	NA	12	16.7	43	59.7	17	18.9
	Total	72	100.0	72	100.0	90	100.0
White/Asian	A	62	86.1	65	90.3	88	97.8
	B	9	12.5	6	8.3	2	2.2
	C	1	1.4	1	1.4	0	0.0
	NA	0	0.0	0	0.0	0	0.0
	Total	72	100.0	72	100.0	90	100.0
White/mixed	A	14	19.4	0	0.0	14	15.6
	B	0	0.0	0	0.0	0	0.0
	C	0	0.0	0	0.0	0	0.0
	NA	58	80.6	72	100.0	76	84.4
	Total	72	100.0	72	100.0	90	100.0
White/other	A	46	63.9	25	34.7	58	64.4
	B	0	0.0	1	1.4	2	2.2
	C	2	2.8	0	0.0	0	0.0
	NA	24	33.3	46	63.9	30	33.3
	Total	72	100.0	72	100.0	90	100.0
White/Wthld. Inf.	A	0	0.0	0	0.0	0	0.0
	B	0	0.0	0	0.0	0	0.0

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent
	C	0	0.0	0	0.0	0	0.0
	NA	72	100.0	72	100.0	90	100.0
	Total	72	100.0	72	100.0	90	100.0

*NA: Insufficient data to compute MH D-DIF