UK Clinical Aptitude Test (UKCAT) Consortium
UKCAT Examination
Technical Report
Executive Summary
Testing Interval: 5 July 2011 – 7 October 2011

Prepared by:
Pearson VUE
March, 2012

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 UK Clinical Aptitude Test Consortium was formed by various medical and dental schools of higher-education institutions in the United Kingdom. The purpose of the UKCAT examination is to help select and/or identify more accurately those individuals with the innate ability to develop professional skills and competencies required to be a good clinician. The test results are to be used by institutions of higher education as part of the process of determining which applicants are to be accepted into the programmes for which they have applied and by the Consortium for research to improve educational services. The goals of the Consortium are to use the UKCAT to widen access for students who desire to study Medicine and Dentistry at the university and to admit those candidates who will become the very best doctors and dentists of the future.

The UKCAT examination was first administered in July 2006 through the Pearson VUE Test Delivery System in testing centers in the United Kingdom and other countries. The 2011 testing period began on 5 July and ended on 7 October. During this period, a total of 24,951 exams were administered. Three forms each of the VR, QR, and AR subtests were used; two forms of the DA subtest were used. The forms were developed from the operational items used in the 2006–2010 administrations and also from items that had been recently tested (2010). All items (operational and pretest) used from 2006 through 2010 were analysed, and those with acceptable item statistics were saved as the active item bank. Items in the active item bank were used to create six versions or forms of the 2011 UKCAT (3 VR/QR/AR * 2 DA). Each candidate was randomly assigned one of the six operational (scored) versions of the cognitive tests and a set of pretest (unscored) items.

Until 2010, the UKCAT analyses—which include item calibration, scaling, and equating—were performed based on a constrained 3-parameter Item Response Theory (3PL-IRT) model. The 3PL-IRT model was chosen in 2006 because of its statistical fitness. The initial scale was established during the 2006 testing window. Subsequent scales were linked back to that reference-group scale. Since 2006, items were calibrated and linked to the reference scale at the end of each test window. Newly calibrated item parameters were used at the test-construction stage to create raw-to-scale-score conversions that would permit immediate scoring for examinees after the end of the testing period. Candidates received four scale scores, one for each of the four subtests. Each cognitive subtest scale score ranges from 300 to 900 with a mean set to 600 in the reference year (2006). For each student, universities received the four subtest scale scores and a total score, which was computed as a simple sum of the four subtest scale scores.

While the 3PL-IRT model has shown good model fit to the data since 2006, it requires a fairly large number of samples for reliable parameter estimation. This practice significantly reduced the number of items that could be pretested each year. To increase the number of pretest items and further strengthen the item bank, Pearson proposed a more parsimonious measurement model such as the Rasch model, which requires a smaller sample to attain reliable parameter estimation. Calibration of the 2006–2010 data showed satisfactory item fit to the Rasch model. More importantly, the Rasch model will allow for up to 2 times more pretest items compared to the 3PL-IRT model. For this reason, all current active items in the bank (including those which appeared in 2011 tests) were rescaled based on the Rasch model at the end of the 2011 test window.

Design of Exam

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

The 2011 exam contains four cognitive subtests: Verbal Reasoning (VR), Quantitative Reasoning (QR), Abstract Reasoning (AR), and Decision Analysis (DA). The VR, QR, and AR subtests contain both operational (scored) and pretest (unscored) items. The DA subtest includes only operational items.

Regular candidates are given 93 minutes to answer a total of 171 items from the VR, QR, AR, and DA subtests. Candidates with special educational needs (SEN) were allotted 117 minutes for the entire exam. The design of the exam is shown below.

Prior to taking the UKCAT examination, candidates are provided access to the UKCAT website for detailed instructions and examples for all subtests.

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. SEN candidates are allotted 26 minutes plus 2 minutes of instruction time.

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 36 items. In addition, candidates are allotted one minute to read general instructions for the subtest. SEN candidates are allotted 27 minutes plus 2 minutes of instruction time.

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 or 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. SEN candidates are allotted 18 minutes plus 2 minutes of instruction time.

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 26 items are scored. Candidates are allowed 31 minutes to answer the 26 items. In addition, candidates are allotted one minute to read general instructions for the subtest. SEN candidates are allotted 38 minutes plus 2 minutes of instruction time.

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). All 26 items within the testlet are presented in a prespecified 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 2011 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, DA, and AR stay fairly close to the previous years (2006-2010).

The average QR score increased significantly from 2009 to 2010 due to the new test conditions, i.e., a shorter test form and longer test time. All items were therefore rescaled in 2010 based on the new test conditions. The new scale was applied to the 2011 QR test and, as a result, the average QR scale score in 2011 returned to the normal range as in 2007, 2008 and 2009.

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.

The 2011 report also includes the performance analysis by candidates' language. Subtest and total scale scores were summarised by candidates' most fluent language and mother tongue. The results indicated that candidates whose most fluent language or mother tongue were English performed significantly better on VR, QR, and DA than candidates who listed other languages as most fluent or their mother tongue. The difference in AR was also observed, but it was less significant than the other three subtests.

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, point-biserial correlations (index 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 2 provides the raw score means, standard deviations, ranges, internal consistency reliabilities (Cronbach's alpha), and standard errors of measurement for each form of each subtest. The means raw scores differences across forms were within 2 points for each subtest. The highest raw score reliabilities were found for AR. This fact can be attributed to the test length. Reliabilities ranged from .69 to .71 for the three VR forms; from .77 to .80 for QR; from .84 to .85 for AR; and .66 and .68 for DA. Standard errors of measurement were on the raw score metric and were approximately 2.9 for VR (number of items = 40),

approximately 2.6 for QR (number of items = 32), 3.5 for AR (number of items = 60), and approximately 2.3 for DA (number of items = 26). The score reliability pattern in 2011 showed slight improvement compared to previous years (2006-2010) and ranged from moderate to high.

Because scale scores are reported to candidates, scale score reliabilities and standard errors are also provided. Table 3a contains the scale score reliabilities and standard errors for each form of the cognitive tests. Unlike the raw score reliability in which 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 are generally good for VR, QR, and AR. Scale score reliabilities improved compared to 2010. Scale score reliabilities ranged from .71 to .74 for the VR forms, from .78 to .81 for the QR forms, from .85 to .87 for the AR forms, and .67 for the two DA forms. Score reliability for AR was higher compared to the other subtests and better reflected the range of reliabilities desired for large-scale testing. Standard errors were approximately 39 for VR, 38 for QR, and 30 for AR. For DA, they averaged around 58. 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 3b 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 forms of the cognitive tests that were administered to a given candidate. 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 3b. The average reliability for total scale score was .89, reflecting good overall reliability. The average standard error was 97.44, which is very reasonable for the range of total scale score.

In summary, score reliabilities of the four cognitive subtests in the 2011 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.

The results of the operational item analyses showed improvements in the overall quality of the 2011 operational pool. Range of difficulty and item discrimination were considerably better in 2011 across the VR, QR, AR and DA subtests.

The pretest statistics, however, were very similar to those of 2010 except the new multiple-choice item type in the VR, which performed better than the conventional 3-option (true, false, cannot tell) item type in terms of discrimination power. Pretest items generally perform less well than the operational items. This is mostly because of the first exposure (i.e., not previously tested and screened) and the smaller sample collected. However, pretest statistics usually improve as they are operationalised and reanalysed based on much larger samples. Pretest item statistics were used not only for screening, but also item bank management (i.e., determining what items would stay in the bank and what items would retire). They were reviewed carefully and provided to item developers for the improvement of future item writing. The 2011 pretest item review meeting was held in February, 2012. In addition, new pretest items were developed to comply with the improved guidelines. The new pretest items will be trialled in the 2012 administration and included in the new active item pool for future test construction.

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. 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 examination and provides the results for the 2011 administration.

Detection of DIF

There are a number of procedures that can be used to detect DIF. 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 to various levels of ability. For the UKCAT examination, 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 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 examination, 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 Categories A and B are not reviewed because of the minor statistical significance. The principal interpretation of Category C items is that—based on the present samples—items flagged in this category 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 social-economic status. Age was separated into groups less than 20 years old and greater than 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 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 4 and 5 show the quantity and percentages of items classified into each of the three DIF categories along with the quantities for which insufficient data were available to compute DIF (Category NA). The results for the operational items are given in Table 4. Those for the pretest items are in Table 5.

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, between white and other race, between white and those who withheld information, and between SEC classes did not meet minimal sample size requirements. The percent of items not meeting sample size requirements ranged from 0% to 7%. In most comparisons, only a small amount (less than 2%) did not meet the sample requirement. These items failed to meet the minimal sample requirement due to the

relatively small samples collected in the focal groups (e.g., age > 35 and ethnic information withheld). These items will be reevaluated for DIF when they are used in future operational forms.

For the operational pools, there were 17 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 0.3%. Of these 17 occurrences, 8 occurred in the Age <20/>35 comparison, 3 in the White/Black comparison, 1 in the White/Asian comparison, 3 in White/Other, 1 in White/Withheld comparison, and 1 in SEC 1/2 comparison. No other comparison groups showed signs of significant DIF. For the pretest items, there were 16 occurrences of Category C DIF, a number that was less than .4% of all comparisons. 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). *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 Group

Test	Total N	Mean	Standard Deviation	Minimum	Maximum
Verbal Reasoning	24951	580.48	74.35	300	890
Quantitative Reasoning	24951	630.69	84.04	300	900
Abstract Reasoning	24951	624.91	80.87	300	900
Decision Analysis	24951	639.64	100.54	300	900
Total Scale Score	24951	2475.72	263.23	1320	3360

Table 2: Raw Score Test Statistics

Test	Form	N Items	N Candidates	Mean	SD	Min	Max	Alpha	SEM
Verbal Reasoning	1	40	8728	23.18	5.14	2	39	0.69	2.87
	2	40	8050	23.30	5.40	5	39	0.69	2.99
	3	40	8173	23.59	5.49	3	39	0.71	2.97
Quantitative Reasoning	1	32	8728	15.43	5.58	0	32	0.77	2.66
	2	32	8050	15.94	5.81	1	32	0.80	2.62
	3	32	8173	16.30	5.45	1	32	0.77	2.61
Abstract Reasoning	1	60	8728	38.95	8.76	0	60	0.84	3.51
	2	60	8050	40.57	8.47	5	60	0.84	3.43
	3	60	8173	39.12	9.07	0	59	0.85	3.51
Decision Analysis	1	26	12877	15.45	4.12	0	26	0.68	2.33
	2	26	12074	17.42	3.79	2	26	0.66	2.22

Table 3a: 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	8728	578.22	72.98	300	890	0.71	39.30
	2	40	8050	580.72	75.66	300	890	0.73	39.31
	3	40	8173	582.67	74.43	300	880	0.74	37.95
Quantitative Reasoning	1	32	8728	625.30	82.63	300	900	0.78	38.76
	2	32	8050	629.63	89.48	320	900	0.81	39.00
	3	32	8173	637.49	79.43	350	900	0.79	36.40
Abstract Reasoning	1	60	8728	623.26	80.23	300	900	0.85	30.87
	2	60	8050	630.38	80.16	300	900	0.87	28.90
	3	60	8173	621.29	81.97	300	890	0.86	30.67
Decision Analysis	1	26	12877	637.68	102.49	300	900	0.67	58.88
	2	26	12074	641.73	98.37	300	900	0.67	56.51

Table 3b: Scale Score Reliability and Standard Error of Measurement for Total Scale Score

Reliability		SEM	
Range[a]	Mean	Range	Mean
.87 - .91	.89	91.38 – 103.49	97.44

[a] Based on 6 combinations of cognitive test forms.

Table 4: 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	120	100.00%	91	98.91%	155	100.00%	52	100.00%
	B	0	0.00%	1	1.09%	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%	92	100.00%	155	100.00%	52	100.00%
Age <20/>35	A	103	85.83%	83	90.22%	146	94.19%	46	88.46%
	B	13	10.83%	7	7.61%	8	5.16%	5	9.62%
	C	4	3.33%	2	2.17%	1	0.65%	1	1.92%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	92	100.00%	155	100.00%	52	100.00%
White/Black	A	114	95.00%	85	92.39%	154	99.35%	52	100.00%
	B	6	5.00%	5	5.43%	0	0.00%	0	0.00%
	C	0	0.00%	2	2.17%	1	0.65%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	92	100.00%	155	100.00%	52	100.00%
White/Asian	A	119	99.17%	90	97.83%	155	100.00%	51	98.08%
	B	1	0.83%	1	1.09%	0	0.00%	1	1.92%
	C	0	0.00%	1	1.09%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	92	100.00%	155	100.00%	52	100.00%
White/mixed	A	120	100.00%	92	100.00%	155	100.00%	51	98.08%
	B	0	0.00%	0	0.00%	0	0.00%	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%	92	100.00%	155	100.00%	52	100.00%
White/other	A	117	97.50%	86	93.48%	154	99.35%	49	94.23%
	B	2	1.67%	4	4.35%	1	0.65%	3	5.77%
	C	1	0.83%	2	2.17%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	92	100.00%	155	100.00%	52	100.00%
White/Wthld. Inf.	A	116	96.67%	87	94.57%	153	98.71%	52	100.00%
	B	4	3.33%	4	4.35%	2	1.29%	0	0.00%
	C	0	0.00%	1	1.09%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	92	100.00%	155	100.00%	52	100.00%
SEC Class 1/2	A	120	100.00%	90	97.83%	151	97.42%	52	100.00%
	B	0	0.00%	2	2.17%	3	1.94%	0	0.00%
	C	0	0.00%	0	0.00%	1	0.65%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	92	100.00%	155	100.00%	52	100.00%
SEC Class 1/3	A	120	100.00%	92	100.00%	155	100.00%	52	100.00%
	B	0	0.00%	0	0.00%	0	0.00%	0	0.00%

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Analysis	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent	Count	Percent
	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%	92	100.00%	155	100.00%	52	100.00%
SEC Class 1/4	A	119	99.17%	92	100.00%	154	99.35%	52	100.00%
	B	1	0.83%	0	0.00%	1	0.65%	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%	92	100.00%	155	100.00%	52	100.00%
SEC Class 1/5	A	116	96.67%	91	98.91%	155	100.00%	52	100.00%
	B	4	3.33%	1	1.09%	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%	92	100.00%	155	100.00%	52	100.00%

Note. NA: Insufficient data to compute MH D-DIF

Table 5: DIF Classification. Pretest Pool

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent
Male/Female	A	103	99.04%	142	92.21%	149	99.33%
	B	1	0.96%	12	7.79%	1	0.67%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
Age <20/>35	A	101	97.12%	143	92.86%	144	96.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	3	2.88%	11	7.14%	6	4.00%
	Total	104	100.00%	154	100.00%	150	100.00%
White/Black	A	104	100.00%	150	97.40%	149	99.33%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	3	1.95%	1	0.67%
	NA	0	0.00%	1	0.65%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
White/Asian	A	95	91.35%	139	90.26%	142	94.67%
	B	8	7.69%	14	9.09%	7	4.67%
	C	1	0.96%	1	0.65%	1	0.67%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
White/mixed	A	102	98.08%	153	99.35%	150	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	2	1.92%	0	0.00%	0	0.00%
	NA	0	0.00%	1	0.65%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
White/other	A	102	98.08%	150	97.40%	150	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	2	1.92%	0	0.00%	0	0.00%
	NA	0	0.00%	4	2.60%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
White/Wthld. Inf.	A	103	99.04%	150	97.40%	150	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	1	0.96%	0	0.00%	0	0.00%
	NA	0	0.00%	4	2.60%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
SEC Class 1/2	A	103	99.04%	151	98.05%	149	99.33%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	1	0.67%
	NA	1	0.96%	3	1.95%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
SEC Class 1/3	A	99	95.19%	154	100.00%	146	97.33%
	B	5	4.81%	0	0.00%	3	2.00%
	C	0	0.00%	0	0.00%	1	0.67%

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning	
Comparison Group	MH-DIF Code	Count	Percent	Count	Percent	Count	Percent
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
SEC Class 1/4	A	103	99.04%	152	98.70%	149	99.33%
	B	0	0.00%	0	0.00%	0	0.00%
	C	1	0.96%	0	0.00%	1	0.67%
	NA	0	0.00%	2	1.30%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%
SEC Class 1/5	A	104	100.00%	151	98.05%	150	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%	3	1.95%	0	0.00%
	Total	104	100.00%	154	100.00%	150	100.00%

Note. NA: Insufficient data to compute MH D-D