PEARSON ALWAYS LEARNING

UK Clinical Aptitude Test (UKCAT) Consortium
UKCAT Examination
Cognitive Sections
Technical Report
Executive Summary
Testing Interval: 3 July 2012 – 5 October 2012

Prepared by:
Pearson VUE
December, 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 (UKCAT) 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 2012 testing period began on 3 July and ended on 5 October. During this period, a total of 25,431 exams were administered. Three forms each of the Verbal Reasoning (VR), Quantitative Reasoning (QR), and Abstract Reasoning (AR) subtests were used; two forms of the Decision Analysis (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 pretested in 2011. All items (operational and pretest) used from 2006 through 2011 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 2012 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, Rasch model will allow for up to three times the number of pretest items compared to the 3PL-IRT model. For this reason, all items in the bank were rescaled based on the Rasch model at the end of the 2011 test window. Rasch model was also applied in 2012 item calibration. Using Rasch model, the number of VR pretest items increased from 104 in 2011 to 332 in 2012. QR pretest items increased from 154 to 332, and AR pretest items increased from 150 to 415. This practice effectively strengthened the active item bank.

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 2012 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) are allotted 117 minutes for the entire exam.

Prior to taking the UKCAT exam, 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 sequenced for presentation to candidates. The four items within each testlet are also randomly sequenced 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 sequenced for presentation to candidates. The four items within each testlet are also randomly sequenced 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 sequenced for presentation to candidates. The five items within each set are also randomly sequenced 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

Students' scale scores are reported for each subtest and are based on all scored items in each subtest. The score ranges from 300 to 900 with a mean set to 600 in the 2006 reference sample. Universities receive the subtest scaled scores for each student plus a total score that is a simple sum of the four subtest scores and has a range of 1,200 to 3,600. An IRT calibration model and IRT true-score equating methods were used to transform the raw scores from each form into a common reporting scale.

Table 1 presents summary statistics for each of the subtests plus the total summed scale score for the total group. A total of 25,431 candidate scores were collected during the 2012 testing window and were used in these analyses. The scale-score means varied across the four subtests. The distributions are generally symmetric around their means and reasonably well spread out.

Most of the scale-score results from 2012 paralleled those from 2011. Total group scale-score means were very close for VR, AR, and DA. For QR, changes in test conditions were implemented in 2010, shifting the average QR score up to 673. The 2011 and 2012 QR tests were scaled based on the new bench and the average score restored to be more consistent with the previous years (2006–2009).

The differential patterns of group performance for gender, age, and NS-SEC in 2012 mirrored those from 2006 to 2011. The results for ethnic group high and low values were also nearly the same as previous years.

3.0 TEST AND ITEM ANALYSIS

Test analysis for the operational forms included computation of the raw-score means, standard deviations, internal consistency reliabilities, and standard errors of measurement of each form of each subtest. Similar test analyses were performed and reported for the scale scores.

Item analysis included a complete classical analysis of item characteristics including *p* values and point biserial (indices of item discrimination). IRT analyses included estimation of item difficulty parameter. The IRT parameter estimates in this report were rescaled to the 2006 reference group.

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 mean raw score 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 .74 to .75 for the three VR forms; from .78 to .79 for QR; from .83 to .85 for AR; and .66 and .67 for DA. Standard error of measurement was based on on the raw score metric and was approximately 2.9 for VR (number of items = 40), approximately 2.5 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 2012 showed slight improvement in VR compared to previous years (2006-2010). All reliability indices 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 indicated that scale-score reliabilities were generally good for VR, QR, and AR. Scale-score reliabilities were very similar to those of 2011 except that improvement in VR was observed. Scale-score reliabilities ranged from .78 to .79 for the VR forms. The reliabilities ranged from .71 to .72 in 2011. For QR, reliabilities ranged from .80 to .82. Reliabilities ranged from .86 to .88 for the AR forms, and .67 for the two DA forms. Score reliability for AR was higher than for the other subtests and better reflected the range of reliabilities desired for large-scale testing. Standard errors were approximately 41 for VR, 38 for QR, and 28 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 six combinations of cognitive test forms and, therefore, there were six 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 .90, reflecting good overall reliability. The average standard error was 96.03, which is very reasonable for the range of total scale score.

In summary, score reliabilities of the four cognitive subtests in the 2012 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 introduction of new item type. A strong item bank provides 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 item analyses differed from the 2011 results in the overall quality of the operational pool. Difficulty range and item discrimination were considerably better in 2012 across the VR, QR, and AR subtests. Significant improvement of the pretest statistics and success rates were also observed in 2012. While pretest items generally had poorer statistics than operational items due to the much smaller sample sizes, the pretest success rate increased from the average of 70% in 2011 to 86% in 2012. Note that pretest statistics may change as they are operationalised and reanalysed based on much larger samples. The improvement of the overall pretest item quality is a result of the Item Review Panel and updated item writing guidelines. The practices will be continued in 2013. Several item-writing workshops will be arranged, and new pretest items will be developed according the improved guidelines. These items will be trialled in the 2013 administration.

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 2012 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, an 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 D-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 200 total (focal plus reference) candidate responses. Because significantly more items were pretested in 2012 and pretest items were 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 (e.g., between White and mixed race, other ethnic minorities, and those who withheld information).

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, comparisons between age groups, between white and mixed race, between white and other race, and between white and those who withheld information did not meet minimal sample size requirements. These comparisons failed to meet the minimal sample requirements

due to the relatively small samples 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 3 occurrences of Category C DIF across all cognitive subtests and comparisons. The proportion of Category C DIF out of all possible comparisons across the four cognitive tests was extremely low (i.e., less than 0.07%). Of these 3 occurrences, 1 occurred in the Age <20 / >35 comparison, 1 in the White/Black comparison, and 1 in the White/Asian comparison. No other comparisons showed signs of significant DIF. For the pretest items, there were 12 occurrences of Category C DIF, a number that was less than .2% of all comparisons. Taken together, the results indicated 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	25431	579.65	89.88	300	900
Quantitative Reasoning	25431	656.35	91.28	300	900
Abstract Reasoning	25431	633.09	80.10	300	900
Decision Analysis	25431	646.47	102.23	300	900
Total Scale Score	25431	2515.56	280.20	1270	3460

Table 2: Raw Score Test Statistics

Test	Form	N Items	N Candidates	Mean	SD	Min	Max	Alpha	SEM
Verbal Reasoning	1	40	8176	25.65	5.63	3	39	0.74	2.85
	2	40	9000	26.58	5.62	2	40	0.74	2.87
	3	40	8255	26.13	5.89	4	40	0.75	2.96
Quantitative Reasoning	1	32	8176	15.27	5.18	0	32	0.78	2.46
	2	32	9000	15.67	5.35	0	32	0.79	2.47
	3	32	8255	15.95	5.37	0	32	0.79	2.47
Abstract Reasoning	1	60	8176	38.48	8.10	0	60	0.83	3.35
	2	60	9000	38.86	8.39	3	60	0.84	3.38
	3	60	8255	38.50	8.78	5	59	0.85	3.39
Decision Analysis	1	26	12293	15.79	4.05	0	26	0.67	2.31
	2	26	13138	16.84	3.84	0	26	0.66	2.23

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	8176	573.20	87.39	300	880	0.78	40.99
	2	40	9000	587.79	91.47	300	900	0.79	41.92
	3	40	8255	577.18	89.90	300	900	0.78	42.17
Quantitative Reasoning	1	32	8176	648.94	90.23	300	900	0.80	40.35
	2	32	9000	656.71	93.42	300	900	0.82	39.64
	3	32	8255	663.30	89.36	300	900	0.82	37.91
Abstract Reasoning	1	60	8176	628.15	75.00	300	900	0.86	28.06
	2	60	9000	641.56	81.56	300	900	0.88	28.25
	3	60	8255	628.73	82.61	300	900	0.87	29.79
Decision Analysis	1	26	12293	654.02	108.85	300	900	0.67	62.53
	2	26	13138	639.40	95.08	300	900	0.67	54.62

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

Reliability		SEM	
Range[a]	Mean	Range	Mean
.88 - .92	.90	90.47 – 101.62	96.03

[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 D-DIF Code	Count	Percent	Count	Percent	Count	Percent	Count	Percent
Male/Female	A	119	99.17%	96	100.00%	180	100.00%	52	100.00%
	B	1	0.83%	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	120	100.00%	96	100.00%	180	100.00%	52	100.00%
Age <20/>35	A	120	100.00%	96	100.00%	180	100.00%	46	88.46%
	B	0	0.00%	0	0.00%	0	0.00%	5	9.62%
	C	0	0.00%	0	0.00%	0	0.00%	1	1.92%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	96	100.00%	180	100.00%	52	100.00%
White/Black	A	111	92.50%	91	94.79%	176	97.78%	52	100.00%
	B	9	7.50%	4	4.17%	4	2.22%	0	0.00%
	C	0	0.00%	1	1.04%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	96	100.00%	180	100.00%	52	100.00%
White/Asian	A	120	100.00%	90	93.75%	178	98.89%	51	98.08%
	B	0	0.00%	5	5.21%	2	1.11%	1	1.92%
	C	0	0.00%	1	1.04%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	96	100.00%	180	100.00%	52	100.00%
White/mixed	A	120	100.00%	96	100.00%	180	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%	96	100.00%	180	100.00%	52	100.00%
White/other	A	120	100.00%	96	100.00%	180	100.00%	49	94.23%
	B	0	0.00%	0	0.00%	0	0.00%	3	5.77%
	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%	96	100.00%	180	100.00%	52	100.00%
White/Wthld. Inf.	A	120	100.00%	96	100.00%	180	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	120	100.00%	96	100.00%	180	100.00%	52	100.00%
SEC Class 1/2	A	120	100.00%	96	100.00%	177	98.33%	52	100.00%
	B	0	0.00%	0	0.00%	3	1.67%	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%	96	100.00%	180	100.00%	52	100.00%
SEC Class 1/3	A	120	100.00%	96	100.00%	180	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 D-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%	96	100.00%	180	100.00%	52	100.00%
SEC Class 1/4	A	119	99.17%	95	98.96%	178	98.89%	52	100.00%
	B	1	0.83%	1	1.04%	2	1.11%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	92	100.00%	180	100.00%	52	100.00%
SEC Class 1/5	A	120	100.00%	96	100.00%	178	98.89%	52	100.00%
	B	0	0.00%	0	0.00%	2	1.11%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%	0	0.00%
	Total	120	100.00%	96	100.00%	180	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 D-DIF Code	Count	Percent	Count	Percent	Count	Percent
Male/Female	A	316	95.18%	315	94.88%	407	98.07%
	B	16	4.82%	15	4.52%	7	1.69%
	C	0	0.00%	2	0.60%	1	0.24%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	332	100.00%	332	100.00%	415	100.00%
Age <20/>35	A	0	0.00%	0	0.00%	0	0.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	332	100.00%	332	100.00%	415	100.00%
	Total	332	100.00%	332	100.00%	415	100.00%
White/Black	A	331	99.70%	332	100.00%	414	99.76%
	B	0	0.00%	0	0.00%	0	0.00%
	C	1	0.30%	0	0.00%	1	0.24%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	332	100.00%	332	100.00%	415	100.00%
White/Asian	A	327	98.50%	332	100.00%	412	99.28%
	B	3	0.90%	0	0.00%	1	0.24%
	C	2	0.60%	0	0.00%	2	0.48%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	332	100.00%	332	100.00%	415	100.00%
White/mixed	A	0	0.00%	0	0.00%	0	0.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	332	100.00%	332	100.00%	415	100.00%
	Total	332	100.00%	332	100.00%	415	100.00%
White/other	A	0	0.00%	0	0.00%	0	0.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	332	100.00%	332	100.00%	415	100.00%
	Total	332	100.00%	332	100.00%	415	100.00%
White/Wthld. Inf.	A	0	0.00%	0	0.00%	0	0.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	332	100.00%	332	100.00%	415	100.00%
	Total	332	100.00%	332	100.00%	415	100.00%
SEC Class 1/2	A	332	100.00%	332	100.00%	415	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	332	100.00%	332	100.00%	415	100.00%
SEC Class 1/3	A	332	100.00%	331	99.70%	414	99.76%
	B	0	0.00%	0	0.00%	0	0.00%

		Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning	
Comparison Group	MH D-DIF Code	Count	Percent	Count	Percent	Count	Percent
	C	0	0.00%	1	0.30%	1	0.24%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	332	100.00%	332	100.00%	415	100.00%
SEC Class 1/4	A	332	100.00%	331	99.70%	415	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	1	0.30%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	332	100.00%	332	100.00%	415	100.00%
SEC Class 1/5	A	332	100.00%	332	100.00%	415	100.00%
	B	0	0.00%	0	0.00%	0	0.00%
	C	0	0.00%	0	0.00%	0	0.00%
	NA	0	0.00%	0	0.00%	0	0.00%
	Total	332	100.00%	332	100.00%	415	100.00%

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