UK Clinical Aptitude Test (UKCAT) Consortium
UKCAT Examination

Executive Summary
Testing Interval: 3 July 2017 – 3 October 2017

Prepared by:
Pearson VUE
31 January, 2018

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 agents and 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

List of Tables

Background

Introduction

The UK Clinical Aptitude Test (UKCAT) 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 the professional skills and competencies required to be good clinicians. Institutions of higher education will use the test results to determine which applicants are to be accepted into the courses for which they have applied. The Consortium will also use the test results for research to improve educational services. The goals of the Consortium are to use the UKCAT to widen access for students who wish to study Medicine and Dentistry at university level 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 centres in the United Kingdom and other countries. The 2017 testing period began on 3 July and ended on 3 October. During this period, a total of 24,841 exams were administered. Three forms each of the Verbal Reasoning (VR), Quantitative Reasoning (QR), Abstract Reasoning (AR), Decision Making (DM), and Situational Judgement Test (SJT) subtests were used to generate three UKCAT forms. Each candidate was randomly assigned one of the three operational (scored) versions of the cognitive tests and a set of pretest (unscored) items. In 2017, a new Decision Making subtest was operationalized.

The forms were developed from the operational items used in the 2006 through 2016 (2013 to 2016 for the SJT) administrations and from items that had been pretested during these years. All items (operational and pretest) used from 2006 through 2016 were analysed, and those with acceptable item statistics were saved as the active item bank.

Until 2010, the UKCAT analyses for the cognitive subtests—which included 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 to that reference-group scale. Since 2006, items for the cognitive subtests have been 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 at the end of the testing period. Candidates received four scaled scores, one for each of the four cognitive subtests, in addition to an SJT band. Each cognitive subtest scaled score ranges from 300 to 900, with a mean set to 600 in the reference year (2006 for VR, QR and AR, and 2016 for DM). For each student, universities received an SJT band, four cognitive subtest scaled scores and a total cognitive scaled score, which was computed as a simple sum of the four subtest scaled scores.

While the 3PL-IRT model has been a good model fit to the data since 2006, it requires a fairly large number of samples for reliable parameter estimation. This limitation 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 VUE 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 to 2010 data showed satisfactory item fit to the Rasch model. More importantly, the 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.

The Rasch model was also applied to 2017 item calibration. Using the Rasch model, the number of VR pretest items increased from 104 in 2011 to 234 in 2017, QR pretest items increased from 154 to 234, and AR pretest items increased from 150 to 296. This pretesting practice effectively strengthens the active item bank.

In addition to the cognitive subtests, candidates are awarded one of four bands for the SJT. The SJT was piloted in 2012 and first introduced operationally in 2013 to evaluate non-academic attributes as part of the UKCAT. The purpose of the SJT is to enable the UKCAT to assess a broader range of constructs outside those relating to cognitive ability. SJTs are designed to assess individuals' judgement regarding situations encountered in a target role. Candidates are presented with a set of hypothetical but relevant scenarios and asked to make judgements about possible responses. Candidates' responses are evaluated against a pre-determined scoring key to provide a picture of their situational judgement in that particular context. SJT scenarios are usually based on extensive analysis of the target role, to ensure that test content reflects the most important situations in which to evaluate candidates' judgement, and are concerned with testing non-academic attributes and ethical values rather than knowledge or clinical skills.

Following the pilot in 2012, a Rasch model was used to equate and scale the SJT in 2013. However, following a review of the 2013 test it was determined that the SJT construct is most likely multi-dimensional, rather than uni-dimensional, and thus the Rasch model is not appropriate. Therefore, from 2014, a classical pre-equating approach (Gibson & Weiner, 1998) was used. The equating and scaling approach used for the SJT is based on a different measurement model to the cognitive subtests and thus no comparison can be made between the item statistics or scaled scores calculated for the SJT and the cognitive subtests.

Design of Exam

The UKCAT is an aptitude exam designed to measure innate cognitive abilities in addition to individuals' judgements regarding situations encountered in a target role. It does not measure student achievement and therefore it does not contain any curriculum or science content.

The 2017 exam contained one SJT subtest and four scored cognitive subtests: VR, QR AR, and DM. All sections contained both operational (scored) and pretest (unscored) items. Candidates were given 120 minutes to answer a total of 232 items from the five subtests. Candidates with special educational needs (SEN) are allotted 150 minutes (UKCATSEN) or 180 minutes (UKCAT50) based on UKCAT's pre-approval, and

candidates with special accommodation (UKCATSA) are allotted 120 minutes for the entire exam with flexible breaks. Prior to taking the UKCAT exam, candidates had access to the UKCAT website for detailed instructions and examples from all subtests. The design of the exam is shown in Table 1.

Table 1. UKCAT Exam Design

Subtest	Scored Items	Unscored Items	Total Number of Items	Test Time
VR	10 testlets of 4 items	1 testlet of 4 items	44	21 minutes allowed on items and 1 minute for instruction
QR	8 testlets of 4 items	1 testlet of 4 items	36	24 minutes allowed on items and 1 minute for instruction
AR	10 testlets of 5 items	1 testlet of 5 items	55	13 minutes allowed on items and 1 minute for instruction
DM	1 testlet of 26 items	3 items	29	31 minutes allowed on items and 1 minute for instruction
SJT	20 testlets of 2 to 5 items	1 testlet of 5 items	68	26 minutes allowed on items and 1 minute for instruction

Examinee Performance

Cognitive Subtests

Students' scaled scores are reported for each of the four cognitive subtests and are based on all scored items in each subtest. The cognitive subtest scaled scores range from 300 to 900 with the mean set to 600 in the 2006 reference sample for VR, QR and AR, and in the 2016 reference sample for DM (see Appendix A for update to VR). Universities receive the subtest scaled scores for each student and a total scaled score that is a simple sum of the four subtest scaled scores and has a range of 900 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 2 presents summary statistics for each of the cognitive subtests and the total summed scaled score for the total group. There were 24,841 candidate scores collected during the 2017 testing window. The means for the scaled score varied across the four cognitive subtests.

Table 2. Cognitive Subtest and Total Scaled Score Summary Statistics: Total Group

Test	Total *N*	Mean	*SD*	Min	Max
VR	24,841	569.77	79.02	300	900
QR	24,841	694.56	98.03	300	900
AR	24,841	628.84	85.81	300	900
DM	24,841	647.01	55.92	300	890
Total	24,841	2540.19	250.32	1,220	3,460

The differential patterns of group performance for gender, age, and NS-SEC in 2017 mirrored those from 2006 to 2015. The results for ethnic group high and low values were also similar to previous years.

Situational Judgement Test

Candidates are awarded one of four bands to reflect their performance on the operational items in the SJT.

A classical pre-equating model was used to transform the raw scores from each form onto a common reporting scale, or scaled score. The band that each candidate receives is determined using the scaled score calculated for each candidate. The narrative that accompanies each band is summarised in Table 3.

Table 3. SJT Band Scaled Score Range and Description

Bands	Scaled Score Range	Narrative
Band 1	654–900	Those in Band 1 demonstrated an excellent level of performance, showing similar judgement in most cases to the panel of experts.
Band 2	583–653	Those in Band 2 demonstrated a good, solid level of performance, showing appropriate judgement frequently, with many responses matching model answers.
Band 3	513–582	Those in Band 3 demonstrated a modest level of performance, with appropriate judgement shown for some questions and substantial differences from ideal responses for others.
Band 4	300–512	The performance of those in Band 4 was low, with judgement tending to differ substantially from ideal responses in many cases.

Table 4 presents the number and percentage of candidates in each band for the 24,841 candidates who took the UKCAT during the 2017 testing window. This is broadly in line with expectations.

Table 4. SJT Band Distribution

SJT Band	Number of Candidates	Percentage of candidates
Band 1	6,970	28.1%
Band 2	10,373	41.8%
Band 3	5,160	20.8%
Band 4	2,338	9.4%
Total	24,841	100.0%

Cognitive Subtests: Test and Item Analysis

Test analysis for the operational forms included computation of the mean and standard deviation of the raw score, internal consistency reliabilities, and standard errors of measurement (*SEM*) of each form of each subtest. Similar test analyses were performed and reported for the scaled 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 the item-difficulty parameter based on the Rasch Model with all operational item parameters anchored to benchmark values. This process ensures that newly developed items (pretest items) are on the same scale as the operational items.

Cognitive Subtests: Test Analysis

Table 5 provides the mean and standard deviation of the raw score, internal consistency reliabilities (Cronbach's alpha), and *SEM* for each form of each subtest. The mean raw score differences across forms were within two points for each subtest.

The highest raw score reliabilities were found for AR. This can be attributed to the test length as AR has the largest number of items; generally, reliability increases with test length. Reliabilities ranged from 0.76 to 0.78 for the three VR forms; from 0.80 to 0.81 for QR; 0.83 for the AR forms; and 0.62 to 0.71 for the DM forms. The *SEM* was based 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), approximately 3.1 for AR (number of items = 50), and approximately 2.8 for DM (number of items = 26). The score reliability pattern in 2017 is similar to the 2016 exam. All reliability indices are in line with expectations for comparable tests of this type and length.

Table 5. Raw Score Test Statistics

Test	Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Alpha	*SEM*
VR	1	40	8,948	21.25	6.05	1	39	0.77	2.90
	2	40	7,918	22.84	5.92	0	39	0.76	2.90
	3	40	7,975	22.38	6.15	4	40	0.78	2.88
QR	1	32	8,948	18.23	5.73	0	32	0.81	2.50
	2	32	7,918	17.59	5.66	0	32	0.80	2.53
	3	32	7,975	18.10	5.83	0	32	0.81	2.54
AR	1	50	8,948	31.39	7.53	0	50	0.83	3.10
	2	50	7,918	29.34	7.64	0	50	0.83	3.15
	3	50	7,975	29.19	7.77	0	50	0.83	3.20
DM	1	26	8,948	17.78	4.59	0	33	0.64	2.75
	2	26	7,918	17.76	4.56	0	31	0.62	2.81
	3	26	7,975	18.37	5.11	0	32	0.71	2.75

Candidates receive a scaled score for each cognitive subtest. Therefore, scaled score reliabilities and standard errors are provided in Table 6. 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 scaled scores depends on the conditional reliability at each scaled score point instead

of on item scores. For this reason, the two reliability indices (Cronbach's alpha and marginal reliability of scaled scores) are not directly comparable.

The results indicated that scaled score reliabilities are in line with expectations given the test lengths for VR, QR, AR, and DM. The VR scaled score reliabilities were similar across forms and ranged from 0.74 to 0.76 in 2017. Reliabilities ranged from 0.78 to 0.79 for the QR forms in 2017, AR scaled score reliability ranged from 0.79 to 0.81, and DM scaled score reliability ranged from 0.61 to 0.69. Score reliability for AR is still higher than for the other subtests, partly due to the longer test length.

Table 6. Scaled Score Reliability and Standard Error of Measurement for Cognitive Subtests

Tests	Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Scaled Score Reliability	*SEM*
VR	1	40	8,948	559.05	77.63	300	880	0.75	38.82
	2	40	7,918	579.07	76.84	300	880	0.74	39.18
	3	40	7,975	572.58	81.28	300	900	0.76	39.82
QR	1	32	8,948	698.89	97.75	300	900	0.78	45.85
	2	32	7,918	688.11	97.58	300	900	0.78	45.77
	3	32	7,975	696.11	98.48	300	900	0.79	45.13
AR	1	50	8,948	643.12	85.76	300	900	0.79	39.30
	2	50	7,918	621.64	84.36	300	900	0.81	36.77
	3	50	7,975	619.96	85.21	300	900	0.80	38.11
DM	1	26	8,948	646.05	53.45	300	890	0.61	33.38
	2	26	7,918	646.09	52.95	300	840	0.61	33.07
	3	26	7,975	649.02	61.21	300	880	0.69	34.08

Table 7 contains the ranges and means of reliabilities and standard errors for the total scaled 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 scaled score is a simple sum (linear composite) of the forms of the cognitive tests that were administered to a given candidate. There were three combinations of cognitive test forms, and therefore there were three estimates of total scaled score reliability and standard error. The range of values and the means are reported in Table 7. The average reliability for total scaled score was 0.89, reflecting good overall reliability. The average standard error was 81.71, which is very reasonable for the range of total scaled score.

Table 7. Scaled Score Reliability and Standard Error of Measurement for Total Scale Score

Reliability		*SEM*	
Range[a]	Mean	Range	Mean
0.89-0.90	0.89	81.50-81.93	81.71

[a]Based on three combinations of cognitive test forms

In summary, score reliabilities of the four cognitive subtests in the 2017 UKCAT are in line with expectations given the test type and length. Reliability for the total scaled score was good and comparable to previous years. Variation in score reliability across the three tests can be partially attributed to the length of subtests. Improvement of score

reliability can be attributed to a stronger item bank as this provides higher flexibility in selecting better-fitted (more discriminative and reasonably challenging) items.

Cognitive Subtests: 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 are comparable to 2016 in terms of the overall quality of the operational pool. The observed range in difficulty and item discrimination are comparable to that observed in 2016 for VR, QR, and AR. While pretest items generally had poorer statistics than operational items due to the much smaller sample sizes, the average 2017 pretest success rate is comparable to that of 2016. Note that pretest statistics may change as items become operational and are re-analysed 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. These practices will continue in 2018. Several item-writing workshops will be arranged, and new pretest items will be developed according to the improved guidelines. These items will be pretested in the 2018 administration.

Cognitive Subtests: 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 cognitive subtests within the UKCAT examination and provides the results for the 2017 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 *comparable* members of the other group did. (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
Black: Black – Black/British – African, Black – Black/British – Caribbean, Black – Black/British Other
Asian: Asian – Asian/British – Bangladeshi, Asian – Asian/British – Indian, Asian – Asian/British – Other Asian, Asian – Asian/British – Pakistani
Chinese: Asian – Asian/British – Chinese
Mixed: Mixed – Mixed – Other, Mixed – White/Asian, Mixed – White/Black African, Mixed – White/Black Caribbean
Other: Other ethnic group

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 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

Table 10 (operational items) and Table 11 (pretest items) in Appendix A show the number 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).

In operational DIF analysis, comparisons between age groups did not meet sample size requirements to compute DIF. For pretest items, comparisons between age groups; White and Other ethnic groups; and NS-SEC Class 1 and the other four NS-SEC groups failed to meet the minimum sample size requirements. These items will be re-evaluated for DIF when they are used in future operational forms.

For the operational pools, there were 12 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. Of these 12 occurrences, one occurred in the Male/Female comparison; three occurred in the Age <20 / >35 comparison; six in the White/Black comparison; and two in the White/Chinese comparison.

For the pretest items, there was one occurrence of Category C DIF in the Male/Female comparison group. It should be noted that as pretest items are seen by fewer

candidates, a significant number of comparisons could not be made due to low sample numbers in the focal groups. Taken together, the results indicated very little DIF occurrence in the UKCAT items.

SJT: Test and Item Analysis

Test analyses for the operational forms included raw score summary statistics, internal consistency reliabilities (Cronbach's alpha), and *SEM* for each form of the SJT. Similar test analyses were performed and reported for the scaled scores. Although the scaled scores are not issued to candidates, they are used to determine the bands awarded to candidates and therefore these summary statistics are presented.

SJT item responses are graded using a partial credit model where candidates are awarded a different number of marks depending on the response they select. Furthermore, the maximum score available varies by items depending on the key with some items having available score points of 0, 1, 3, 4 and others using score points of 0, 1, 2, 3.

The SJT items are analysed using Classical Test Theory because a review of the SJT, following the 2013 test window, showed that an IRT approach is not appropriate. Unlike IRT, Classical Test statistics are sample dependent, meaning that they are calculated based on the actual sample of candidates who respond to each item and are not linked to a common benchmark group. Therefore, the item statistics presented for the SJT are not comparable to those presented for the cognitive subtests due to the different measurement models used.

SJT: Test Analysis

The mean and standard deviation of the raw score, internal consistency reliabilities and *SEM* for each form of the SJT are summarised in Table 8. The test statistics are computed based on all candidates who took the SJT. The maximum number of available score points is 233 for all the three forms in 2017, however, it has varied in previous years. Therefore, the best way to compare the raw score is to compare the mean raw score as a percentage of the maximum available score. Raw score reliabilities for the three SJT forms ranged from 0.82 to 0.85. The reliabilities for all SJT forms are good and comparable to 2016. The *SEM* was based on the raw score metric and ranged from 8.51 to 9.02.

Table 8. SJT Raw Score Test Statistics

Form	*N* Items	*N* Candidates	Mean	*SD*	Min	Max	Mean Percent Raw Score	Alpha	*SEM*
1	63	8,948	165.31	21.26	0	211	71%	0.82	9.02
2	63	7,918	168.16	21.97	0	217	72%	0.85	8.51
3	63	7,975	164.31	21.39	0	212	71%	0.83	8.82

The band that candidates receive for the SJT is based on their SJT scaled score and therefore test statistics for scaled scores are provided in Table 9. The scaled scores are linearly related to the raw scores and therefore the raw score reliability applies equally to the scaled scores. This is in contrast to the cognitive subtests where the scaled scores are a transformation of the IRT ability values.

The average scaled score ranged from 601.59 (Form 3) to 616.00 (Form 2). The maximum possible scaled score varies across the forms (826 on Form 1, 828 on Form 2, and 820 on Form 3). The *SEM* of the scaled scores averaged 28 for the SJT. The differences in average scaled score between the forms are small and all within one *SEM*.

Table 9. SJT Scale Score Test Statistics

Form	*N* Items	*N* Candidates	Mean	SD	Min	Max	*SEM*
1	63	8,948	609.45	67.63	300	756	28.69
2	63	7,918	616.00	71.45	300	776	27.67
3	63	7,975	601.59	67.43	300	753	27.80

SJT: Item Analysis

The SJT items are analysed using Classical Test Theory for both operational and pretest items.

The results of the item analyses are comparable to the 2016 results in the overall quality of the operational pool. Difficulty range and item discrimination in 2017 were also comparable to that observed in 2016. Note that pretest statistics may change as they become operational and are re-analysed based on much larger samples.

SJT: Differential Item Functioning

Introduction

The DIF analysis is a procedure used to determine if test items are fair and appropriate for assessing the ability of various demographic groups. It is based on the assumption that test-takers who have similar ability (based on total test scores) should perform in similar ways on individual test items, regardless of their demographic group. Note that as the measurement model used for the SJT is different to that used for the cognitive sections, the method for identifying DIF is also different.

Detection of DIF

DIF analysis was performed on the items in the pool using a hierarchical regression approach using the equated scaled score. The polytomous scoring of these items makes this approach appropriate. For each comparison, the first column indicates the size of increase in the variance in item responses explained by the regression equation when the group membership variable and an interaction variable of group membership with SJT score was added to the equation.

Criteria for Flagging Items

Effects that explain less than 1% of score variance (R squared change <0.01) are considered negligible for flagging purposes and items that do not reach significance, or explain less than this proportion of variance, are labelled ‘A’ and can be considered free of DIF. Larger effects, where the group variable has a significant beta coefficient are labelled ‘B’ or ‘C’. Changes of 0.01 or above are considered slight to moderate and labelled ‘B’, unless all of the change is explained by the interaction term, in which case they are labelled ‘A’. Changes above 0.05 (5% of the variance in responses) are

considered moderate to large and are labelled ‘C’ where there is a significant main effect of the group difference variable.

With large samples and many comparisons, the probability of Type 1 errors is high and effect sizes that are too small to be of substantive interest may reach statistical significance. Because of the frequency of Type 1 errors with large numbers of comparisons, item flags were used as signals to further review items rather than as indicators that items needed to be dropped from the pool. At the 95% significance level, Type 1 errors would be expected for 5% of the comparisons tested. Therefore, where the number of flags is similar to these figures, it is possible that all the effects are Type 1 errors. In addition, DIF is a necessary but not sufficient condition for bias: bias only exists if the difference is illegitimate, i.e., if both groups should be performing equally well on the item.

All items with moderate to large DIF will be further reviewed and dropped from the operational item pool where any potential unfairness in the content is identified.

Comparison Groups for DIF Analysis

The same UKCAT DIF comparison groups used for the cognitive subtests were used for the SJT.

Sample Size Requirements

If the sample size for the analysis is less than 200, the sample is not large enough to undertake analysis and therefore DIF is not reported.

DIF Results

Table 12 (operational items) and Table 13 (pretest items) in Appendix B, show the number 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 N<200).

In operational DIF analysis, all items met sample size requirements to compute DIF for all comparison groups for the SJT. For some pretest items, comparisons between White and Black, White and Chinese, White and Mixed, between NS-SEC Class 1 and Class 2, and between Class 1 and Class 4 did not meet minimal sample size requirements. These items will be re-evaluated for DIF when they are used in future operational forms.

For the operational SJT pool, there were no occurrences of Category C DIF and 20 instances of Category B DIF. For the pretest items, there were three occurrences of Category C DIF. It should be noted that, as pretest items are seen by fewer candidates, a significant number of comparisons could not be made due to low sample numbers in the focal groups.

Taken together, the results indicated very little DIF occurrence in the UKCAT SJT items.

References

Gibson, W. M., & Weiner, J. A. (1988). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement, 35*, 297-310.

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.

Appendix A: Cognitive Subtest DIF Summary Tables

Table 10. DIF Classification. Operational Pool

Comparison Group	MH D-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Making	
		N Items	Percent	*N* Items	Percent	*N* Items	Percent	*N* Items	Percent
Male/Female	A	117	98%	96	100%	150	100%	78	100%
	B	2	2%	0	0%	0	0%	0	0%
	C	1	1%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%
Age <20/>35	A	78	65%	64	67%	92	61%	48	62%
	B	0	0%	0	0%	0	0%	0	0%
	C	1	1%	0	0%	0	0%	2	3%
	NA	41	34%	32	33%	58	39%	28	36%
	Total	120	100%	96	100%	150	100%	78	100%
White/Black	A	118	98%	96	100%	147	98%	67	86%
	B	1	1%	0	0%	3	2%	6	8%
	C	1	1%	0	0%	0	0%	5	6%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%
White/Asian	A	118	98%	96	100%	149	99%	77	99%
	B	2	2%	0	0%	1	1%	1	1%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%
White/ Chinese	A	120	100%	94	98%	150	100%	78	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	2	2%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%
White/Mixed	A	120	100%	96	100%	150	100%	78	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%
SEC Class 1/2	A	120	100%	96	100%	149	99%	76	97%
	B	0	0%	0	0%	1	1%	2	3%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%
SEC Class 1/3	A	120	100%	96	100%	150	100%	78	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%

Comparison Group	MH D-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Making	
		N Items	Percent	*N* Items	Percent	*N* Items	Percent	*N* Items	Percent
SEC Class 1/4	A	120	100%	96	100%	150	100%	78	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%
SEC Class 1/5	A	120	100%	96	100%	149	99%	77	99%
	B	0	0%	0	0%	1	1%	1	1%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	120	100%	96	100%	150	100%	78	100%

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

Table 11. DIF Classification. Pretest Pool

Comparison Group	MH D-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Making	
		N Items	Percent	*N* Items	Percent	*N* Items	Percent	*N* Items	Percent
Male/Female	A	234	100%	234	100%	295	100%	215	100%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	1	0%	0	0%
	NA	0	0%	0	0%	0	0%	0	0%
	Total	234	100%	234	100%	296	100%	215	100%
Age <20/>35	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	234	100%	234	100%	296	100%	215	100%
	Total	234	100%	234	100%	296	100%	215	100%
White/Black	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	234	100%	234	100%	296	100%	215	100%
	Total	234	100%	234	100%	296	100%	215	100%
White/Asian	A	234	100%	225	96%	296	100%	173	80%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	0	0%	9	4%	0	0%	42	20%
	Total	234	100%	234	100%	296	100%	215	100%
White/ Chinese	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	234	100%	234	100%	296	100%	215	100%
	Total	234	100%	234	100%	296	100%	215	100%
White/Mixed	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	234	100%	234	100%	296	100%	215	100%
	Total	234	100%	234	100%	296	100%	215	100%

Comparison Group	MH D-DIF Code	Verbal Reasoning		Quantitative Reasoning		Abstract Reasoning		Decision Making	
		N Items	Percent	*N* Items	Percent	*N* Items	Percent	*N* Items	Percent
SEC Class 1/2	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	234	100%	234	100%	296	100%	215	100%
	Total	234	100%	234	100%	296	100%	215	100%
SEC Class 1/3	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	234	100%	234	100%	296	100%	215	100%
	Total	234	100%	234	100%	296	100%	215	100%
SEC Class 1/4	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	234	100%	234	100%	296	100%	215	100%
	Total	234	100%	234	100%	296	100%	215	100%
SEC Class 1/5	A	0	0%	0	0%	0	0%	0	0%
	B	0	0%	0	0%	0	0%	0	0%
	C	0	0%	0	0%	0	0%	0	0%
	NA	234	100%	234	100%	296	100%	215	100%
	Total	234	100%	234	100%	296	100%	215	100%

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

Appendix B: SJT DIF Summary Tables

Table 12. SJT DIF Classification: Operational Pool

Comparison Group	Degree of DIF					
	A		B		C	
	N Items	Percent	*N* Items	Percent	*N* Items	Percent
Male/Female	133	99%	1	1%	0	0%
Age <20/>35	129	96%	5	4%	0	0%
White/Black	128	96%	6	4%	0	0%
White/Asian	126	94%	8	6%	0	0%
White/Chinese	134	100%	0	0%	0	0%
White/Mixed	134	100%	0	0%	0	0%
SEC Class 1/2	134	100%	0	0%	0	0%
SEC Class 1/3	134	100%	0	0%	0	0%
SEC Class 1/4	134	100%	0	0%	0	0%
SEC Class 1/5	134	100%	0	0%	0	0%

Table 13. SJT DIF Classification: Pretest Pool

Comparison Group	Degree of DIF							
	A		B		C		N<200	
	N Items	Percent	*N* Items	Percent	*N* Items	Percent	*N* Items	Percent
Male/Female	235	95%	11	4%	1	0%	0	0%
Age <20/>35	231	94%	15	6%	1	0%	0	0%
White/Black	153	62%	7	3%	0	0%	87	35%
White/Asian	234	95%	13	5%	0	0%	0	0%
White/Chinese	60	24%	4	2%	0	0%	183	74%
White/Mixed	86	35%	1	0%	1	0%	159	64%
SEC Class 1/2	243	98%	1	0%	0	0%	3	1%
SEC Class 1/3	242	98%	5	2%	0	0%	0	0%
SEC Class 1/4	243	98%	1	0%	0	0%	3	1%
SEC Class 1/5	245	99%	2	1%	0	0%	0	0%