Understanding the Dimensionality and Reliability of the Cognitive Scales of the UK Clinical Aptitude test (UKCAT): Summary Version of the Report

Dr Paul A. Tiffin, Reader in Psychometric Epidemiology, Durham University

December 2013

Background

Although used to select medical and dental students since 2006 the structure of the UKCAT has not been published to date. The dimensionality of the UKCAT responses has implications for how the test scores are summarised and used in the selection process. Moreover, a number of independent researchers have questioned the very high (internal) reliability figures cited by the annual technical reports. Understanding the dimensionality of the UKCAT will allow for accurate reliability estimates to be derived using widely accepted methods.

Consequently the UKCAT research panel commissioned an item-level analysis involving both exploratory and confirmatory factor analyses in order to elicit the structure of the test and hence establish its reliability. The analysis was conducted in December 2013.

Summary of Methods

The UKCAT is delivered in a number of versions which use differing permutations of the four subtests (*abstract reasoning* [AR], *decision analysis* [DA], *quantitative reasoning* [QR] and *verbal reasoning* [VR]. For this study item responses from the 2013 version 1 were analysed from a total of 5,053 testees, randomly divided into 2,488 observations to be used as the 'training' dataset and 2,565 observations to be used as the 'prediction' or 'confirmatory' dataset. A parallel analysis (Horn 1965) adapted for binary response data was used to evaluate the dimensionality of the response data. In order to reduce the risk of such *item dependency* (which could artefactually inflate reliability coefficients) only one item from each group of questions shown to be related to the same stem was randomly selected for inclusion in the final analysis. The only subtest that this approach could not be used with was *decision analysis* as all the items used in a particular form of *decision analysis* (only two forms are used in each sitting) were related to a single stem.

Once plausible structures for the response data were established using the parallel analysis and exploratory factor analyses, these were then tested using the confirmatory dataset from version 1 of the UKCAT 2013. Confirmation of the data structure was also conducted in version 6 of the UKCAT test conducted in 2013. Thus confirmation was conducted in

datasets which related to both the same version and a completely different version of the UKCAT.

In terms of 'reliability'; addition to McDonald's omega, a binary form of Cronbach's Alpha was derived from FACTOR (Lorenzo-Seva, 2013) as well as a conventional form of Cronbach's Alpha that was derived from STATA version 13. Test information curves for the subtests of the UKCAT were plotted based on the findings and confirmatory factor analyses of responses to version 6 the test from 2013 utilising a two parameter logistic model (2-PL) using Mplus v7.1.

Main findings

The findings from the parallel analysis strongly suggested that no more than three distinct factors could be extracted from the response data in the 2,488 who were in the UKCAT 2013 version 1 training dataset. One, two, three and the current four factor models were all fitted to both the confirmatory dataset for version 1 of the 2013 UKCAT (N=2,565) and all the responses to version 6 of the 2013 UKCAT (N=4,042). Fit between the competing models differed relatively little, as estimated by the Tucker-Lewis Index (TLI- regarded as more conservative and trustworthy as it penalises model complexity(Hu and Bentler 1999)). A secondary, generalised, factor could also be conceptualised, lying beneath the four factors representing the scales. This hierarchical factor model also showed an acceptable fit to the response data (TLI=0.96).

In particularly, it is worth noting that a two factor model with verbal (VR) versus non-verbal (DA, AR and QR-see Figure 1) had a TLI of 0.941 compared to 0.946 for the original, four factor (i.e. four subscale) model. There may be theoretical and empirical reasons why the test should perhaps be regarded as composed of verbal and non-verbal components (Deary et al. 2007).

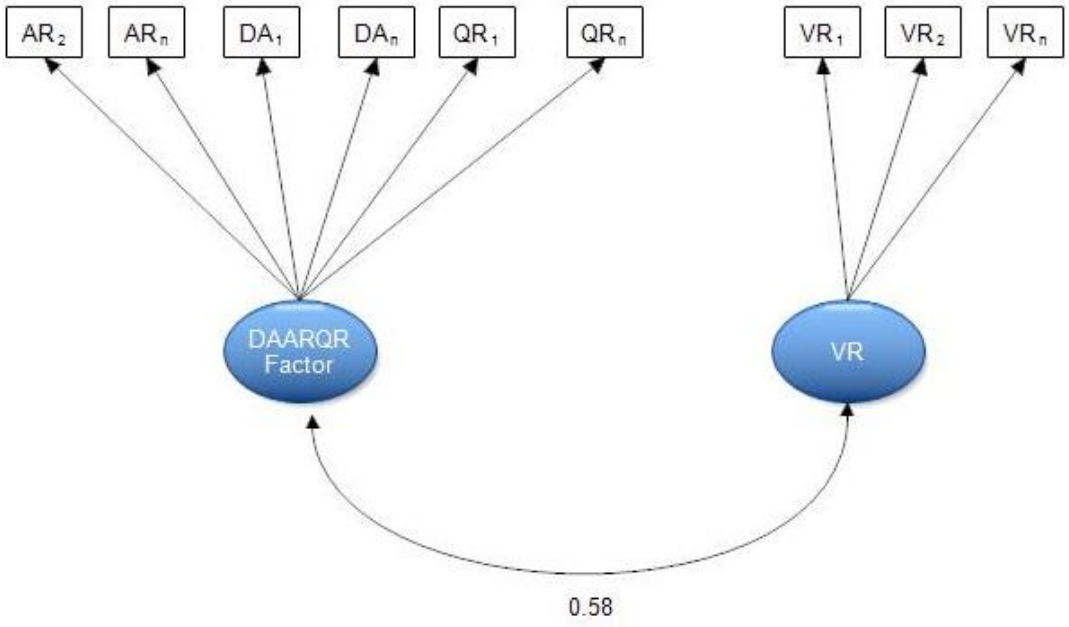

Figure 1. Path diagram depicting correlations between the factors relating to an alternative postulated two-factor structure for the UKCAT response scores. This model is based on a general factor represented by the DA, AR and QR items and second factor represented by the VR items. Note, for simplicity only a few indicators (shown as squares) are included for each latent variable. Error (residual) variances and loadings are also omitted for clarity.

Internal reliability consistency indices were generated for the UKCAT subscales and total scores (in the case of McDonald's omega the latter was generated for 'G' factor scores). Reliability values for the UKCAT cognitive scales were acceptable, though generally lower than those cited in the previous technical reports with omega values of 0.49 (QR) to 0.87 (DA). Cronbach alpha values (binary version) were almost identical (Table 1).

In item response theory (IRT) more importance is placed on *test information* compared to the traditional 'reliability'. Test information is reciprocal to the standard error of measurement (SEM). This is intuitive; the smaller the error the more information is available from a test to discriminate between testees. The test information curves for the UKCAT 2013 test (version 6) are depicted in Figures 2-5. The horizontal axes relate to the level of trait under evaluation (as a z score with mean of zero and sd of 1). It can be seen that the DA and AR subscales yield relatively information on more able candidates- the problem is particularly acute for DA in this form of the test evaluated. In contrast, the information 'max's out' around the average level of ability for the QR and VR subtest.

Subtest	Alpha (range) (PV Technical Report, 2012)	Cronbach's alpha (usual)	Cronbach's alpha (binary)	McDonald's Omega*
AR	0.86-0.87	0.53	0.65	0.66
DA	0.66-0.67	0.69	0.87	0.87
QR	0.78-0.79	0.38	0.49	0.49
VR	0.74-0.75	0.45	0.62	0.62
All items (i.e. loadings on 'G' factor for Ω_h)	NA	0.77	0.89	0.41

Table 1. Internal reliability consistency estimates for the UKCAT scales derived from responses to version 6 of the UKCAT in 2013 (N=4,042).

*Omega is generated as Ω_t for subtests and Ω_h for a general factor; the latter reflecting the internal reliability consistency of a summed score.

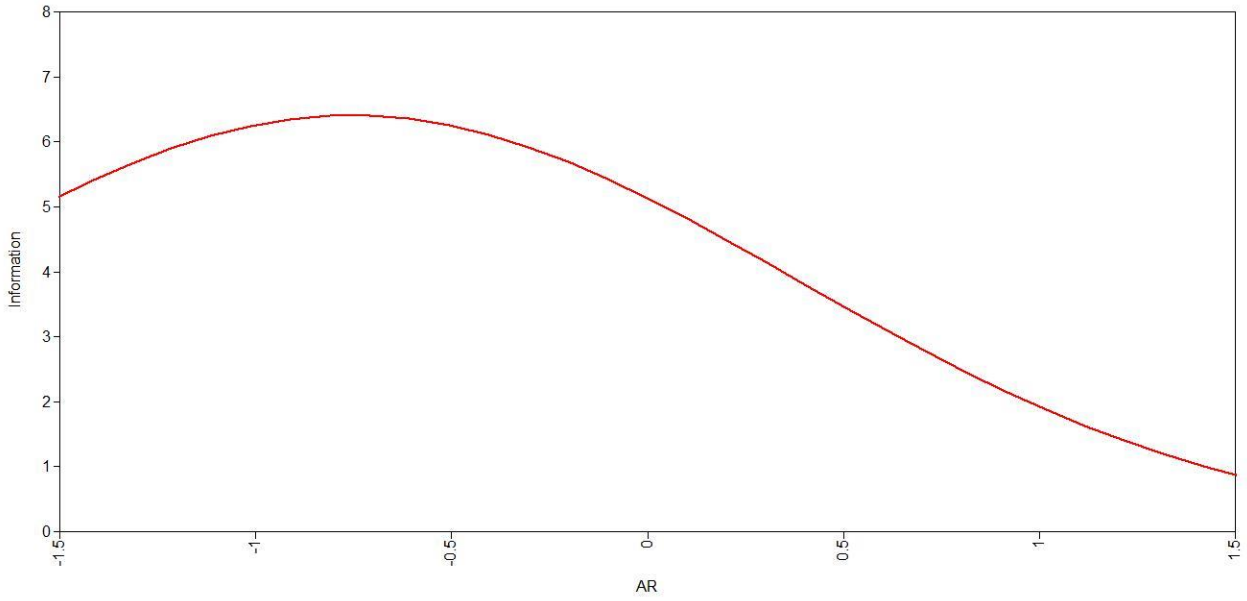

Figure 2. The test information curve for the *abstract reasoning* subtest of the UKCAT. The vertical axis represents the amount of information available on each candidate whilst the horizontal axis represents trait/ability level (theta) for a candidate. Trait is expressed as a standardised score mean of zero and standard deviation of one.

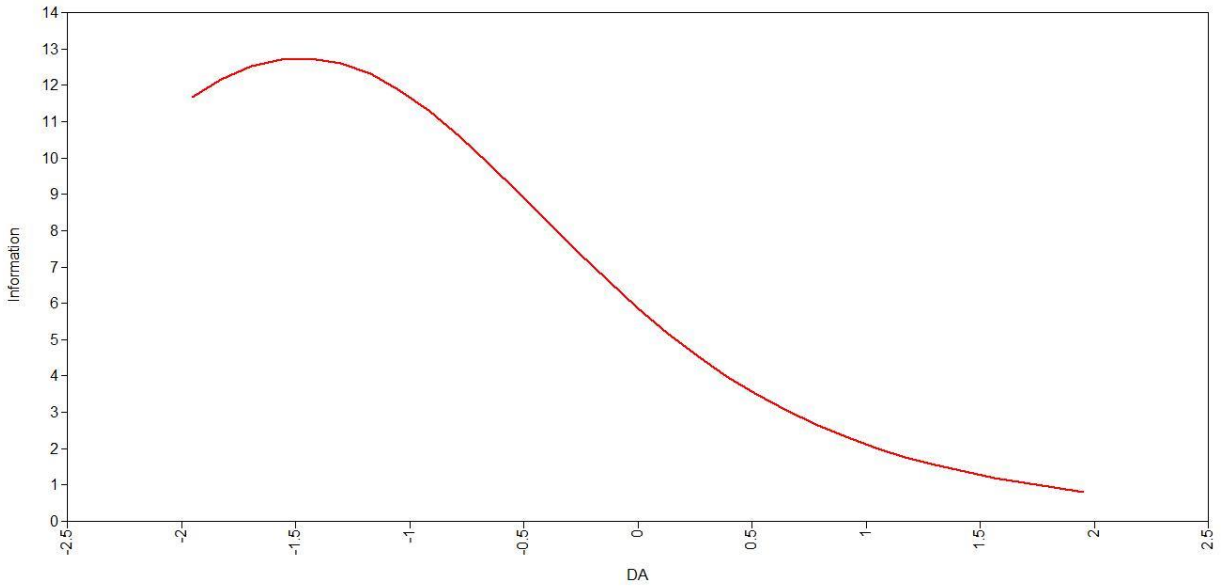

Figure 3. The test information curve for the *decision analysis* subtest of the UKCAT. The vertical axis represents the amount of information available on each candidate whilst the horizontal axis represents trait/ability level (theta) for a candidate. Trait is expressed as a standardised score mean of zero and standard deviation of one.

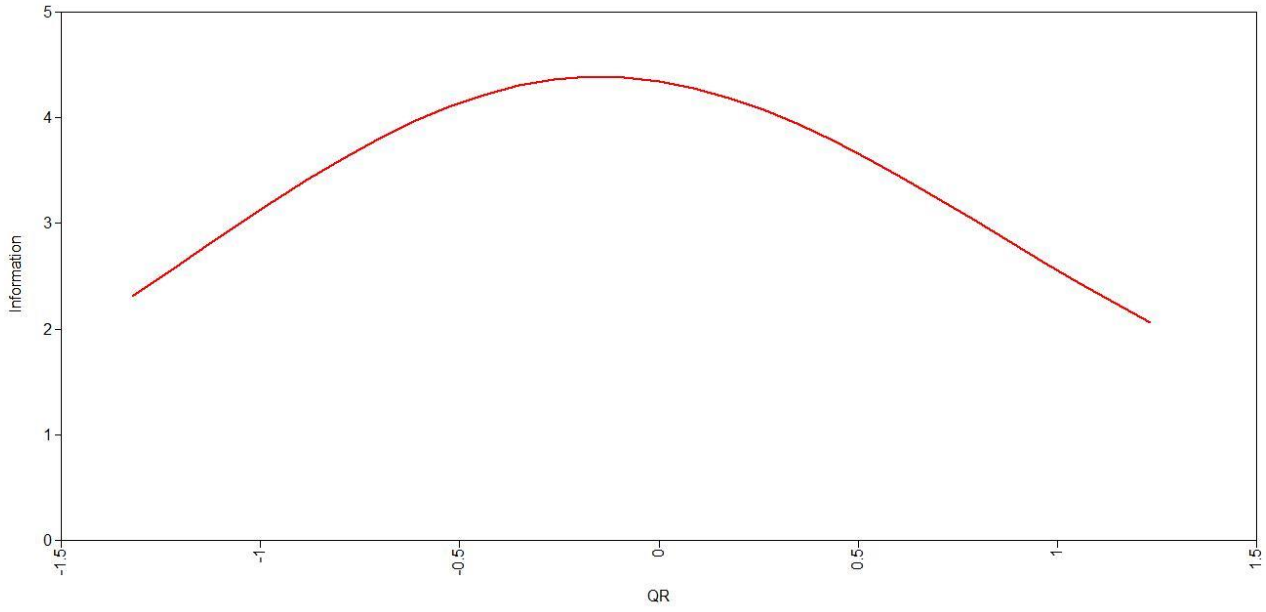

Figure 4. The test information curve for the quantitative reasoning subtest of the UKCAT. The vertical axis represents the amount of information available on each candidate whilst the horizontal axis represents trait/ability level (theta) for a candidate. Trait is expressed as a standardised score mean of zero and standard deviation of one.

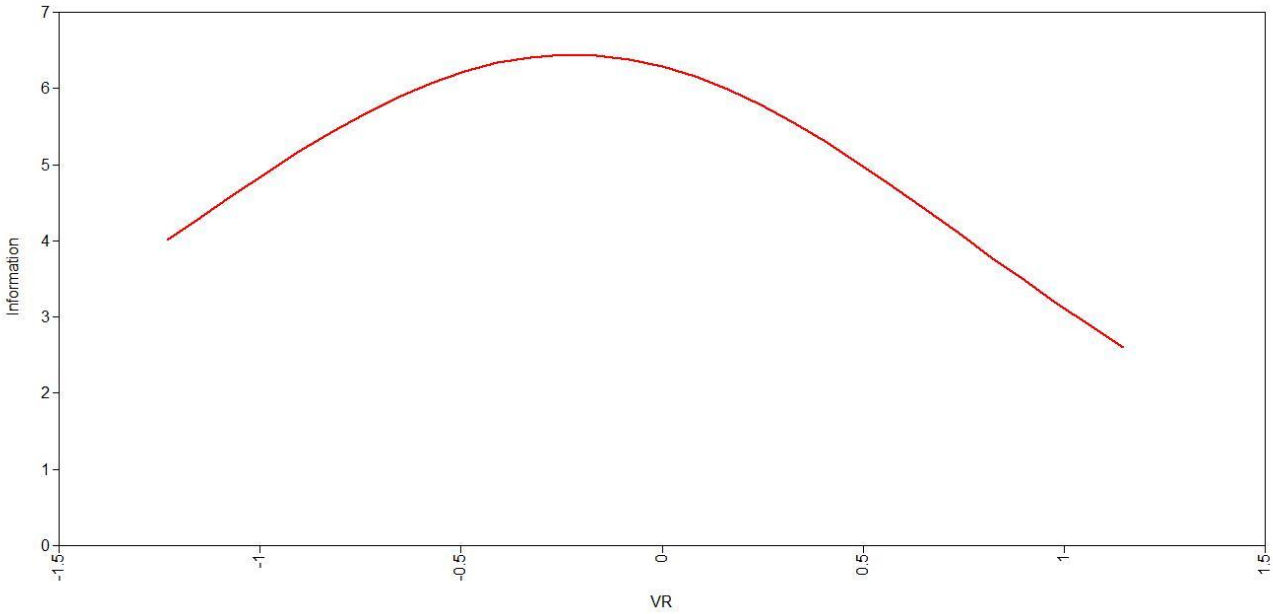

Figure 5. The test information curve for the verbal reasoning subtest of the UKCAT. The vertical axis represents the amount of information available on each candidate whilst the horizontal axis represents trait/ability level (theta) for a candidate. Trait is expressed as a standardised score mean of zero and standard deviation of one.

Summary

The UKCAT responses fit well into a multidimensional model that can be conceptualised as also being related to a general ('G') factor. This finding lends some support to the use of the total UKCAT score in selection. There may also be a case for distinguishing between the verbal and non-verbal scales of the UKCAT in the selection process (e.g. summing the VR score and the average for QR, AR and DA). The internal reliability consistency of the UKCAT is acceptable, though not generally as high as inferred from indices reported in the Pearson Vue technical reports. This may have been due to the dependency of items nested in the same question stem in the original analyses. The *abstract reasoning* and *decision analysis* subtests are mainly composed of items that are relatively easy. This may lead to difficulties reliably distinguishing between more able candidates and these scales may benefit from the inclusion of more difficult items. Given the relatively high degree of correlation between the 'non-verbal' scores of the UKCAT it may be that they are tending to measure the same underlying construct. This may give rise to an opportunity to shorten this aspect of the test, creating space to include novel tests or items tapping into different traits or abilities.

Acknowledgements

Many thanks to Dr Brad Wu (Senior Psychometrician at Pearson Vue) for his invaluable help in obtaining the data and meta-data used in the analyses. I am also grateful to the UKCAT Board for agreeing to fund this work.

References

Deary, I. et al. (2007) Intelligence and educational achievement. Intelligence 35:13-21.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika **30**: 179-185.

Hu, L. and P. M. Bentler (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal **6**(1): 1-55.

Lorenzo-Seva, U. (2013). FACTOR. University of Tarragona.

McDonald, R. P. (1978). Generalizability in factorable domains: domain validity and

generalizability. Educational and Psychological Measurement **38**(1): 75-79.

McManus, I.C. et al. (2013) Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: meta-regression of six UK longitudinal studies. BMC Medicine **11**: 243.

Schmid, J. and J. N. Leiman (1957). The development of hierarchical factor solutions. Psychometrika **22**: 53-61.

Wu, B. (2012). Technical Report UK Clinical Aptitude Test (UKCAT) Consortium. Chicago, IL, Pearson VUE.

Zinbarg, R. E., et al. (2005). Cronbach's α, Revelle's β, and McDonald's Ω_h: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika **70**(1): 123-133.