Understanding the measurement model of the UKCAT Situational Judgment Test: Summary Report

Paul Tiffin & Madeline Carter

May 2015

Background

In order to develop a reliable and defensible test it is important to understand the 'measurement model' represented by responses to the items. In addition, how the test should be used within the selection process should be partly determined by its psychometric characteristics, and in particular where the information on candidates is maximal. For example, relatively easy tests demonstrate most information around the lower end of ability, allowing for discrimination between candidates with lower ability and moderate/high ability. Such tests may prove to be effective 'screeners' but would not be expected to accurately discriminate between more able candidates.

This work seeks to replicate some of the earlier analysis performed on the cognitive subscales of the UKCAT, though with the added complexity of a polytomous, and sometimes weighted, scoring system for responses. For these analyses, data have been provided by the UKCAT Board for test responses from 23,884 UKCAT candidates who had taken six different (though overlapping) forms of the SJTs in 2014 (for entry in 2015).

1. Understanding the dimensionality underlying the test responses

In order to explore the underlying dimensionality of the SJT responses, parallel analyses were conducted, implemented within the freeware software package FACTOR, and adapted for ordinal data. In order to eliminate the possible dependency (i.e. correlated residuals) between item responses related to the same scenario/stem, only one item (selected at random) from each set of three or four questions related to the same stem was included in the parallel analysis. The data for each form was randomly divided into two subsets - one which was 'exploratory' and another which was held back for 'confirmatory' analysis. Also, note that unweighted scorings (i.e. 0, 1, 2, 3 for all responses with '3' as the most desirable) were used so that the thresholds between categories could be inspected. The appropriateness of weighting response scores will be discussed later.

The findings from five of the forms analysed (R1, R2, R4, R5, R6) suggested that only one significant dimension lay behind the response patterns observed, when considering the percentage of variance explained by each successive dimension in the 95th percentile of the random data generated. In all six forms, roughly half to two-thirds of the items tested loaded relatively substantially on the main factor (i.e. loadings close to or exceeding a magnitude of .3). The only exception to the unidimensional structure observed was for form R3 where there was more evidence for a second dimension (see also section 2, below). In this latter form the real data just outperformed the random data in terms of a second dimension explaining a portion of the variance in responses. For this reason the possibility of a two factor model was explored for form R3 (see below).

Findings from Confirmatory and Exploratory Factor Analyses on Responses in Form R1

For the responses to form R1, exploratory and confirmatory factor analytic models (EFA and CFA respectively) were developed (the latter using the held back, 'validation response data set'). Note, for the EFA and CFA models all the items that loaded substantially on the main factor, not just those randomly selected from each stem and used in the parallel analysis, were included. The findings from these analyses suggested that if only items that loaded relatively substantially on the main factor were included as indicators, then an adequately fitting unidimensional factor model could be achieved (i.e. TLI and CFI>.90). However, this required some modification, in that a small number of correlated residuals (around 4 to 6 pairs) had to be allowed between items as there was dependency between responses. In some cases these correlated residuals were between items related to the same stem, but not always. A test information curve suggested, as expected, that most of the information was produced on candidates below the average ability level for 'situational judgement' (see Figure 1).

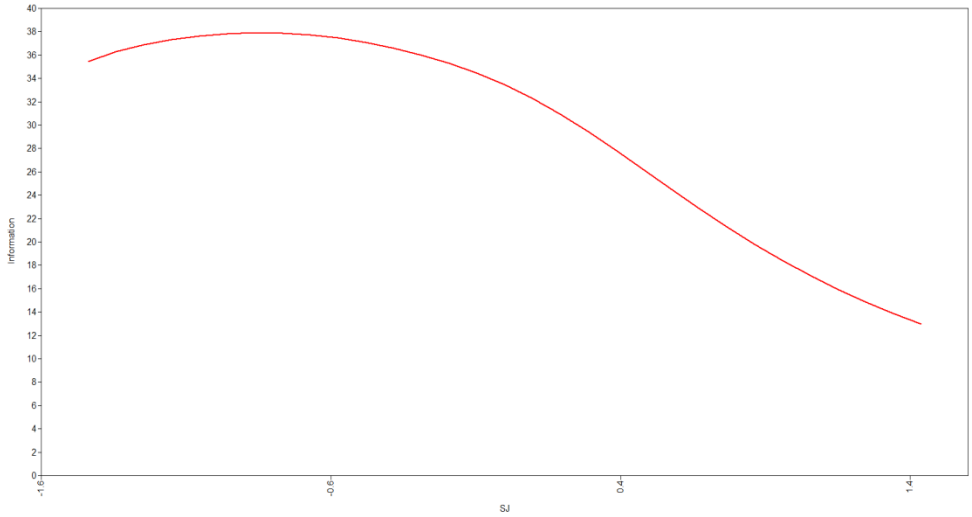


Figure 1. Test information curve (TIC) for form R1. This was derived from a one factor CFA. In order to generate this curve the CFA was reparameterized into a two parameter logistic (2-PL) model, utilising the graded response model. Thus, the curve takes into account both the item difficulties and discrimination values. SJ='Situation Judgement' [trait level]

The possibility of a two factor model was considered for the responses to the items in form R3. The methods of model building were the same as that used in form R1. Only those items that loaded substantially (magnitude >0.3) on one of the two factors were included. Modification indices were used to iteratively guide model modification until a satisfactory fit was achieved (CFI=0.90, TLI=0.90). The modifications required to achieve adequate fit were: four cross-loadings, and; three pairs of correlated residuals were permitted between item pairs that demonstrated marked dependency. The resulting model is depicted in Figure 2.

The reliability indices derived from the unidimensional factor analysis (as implemented in FACTOR) were acceptable and are depicted in Table 1.

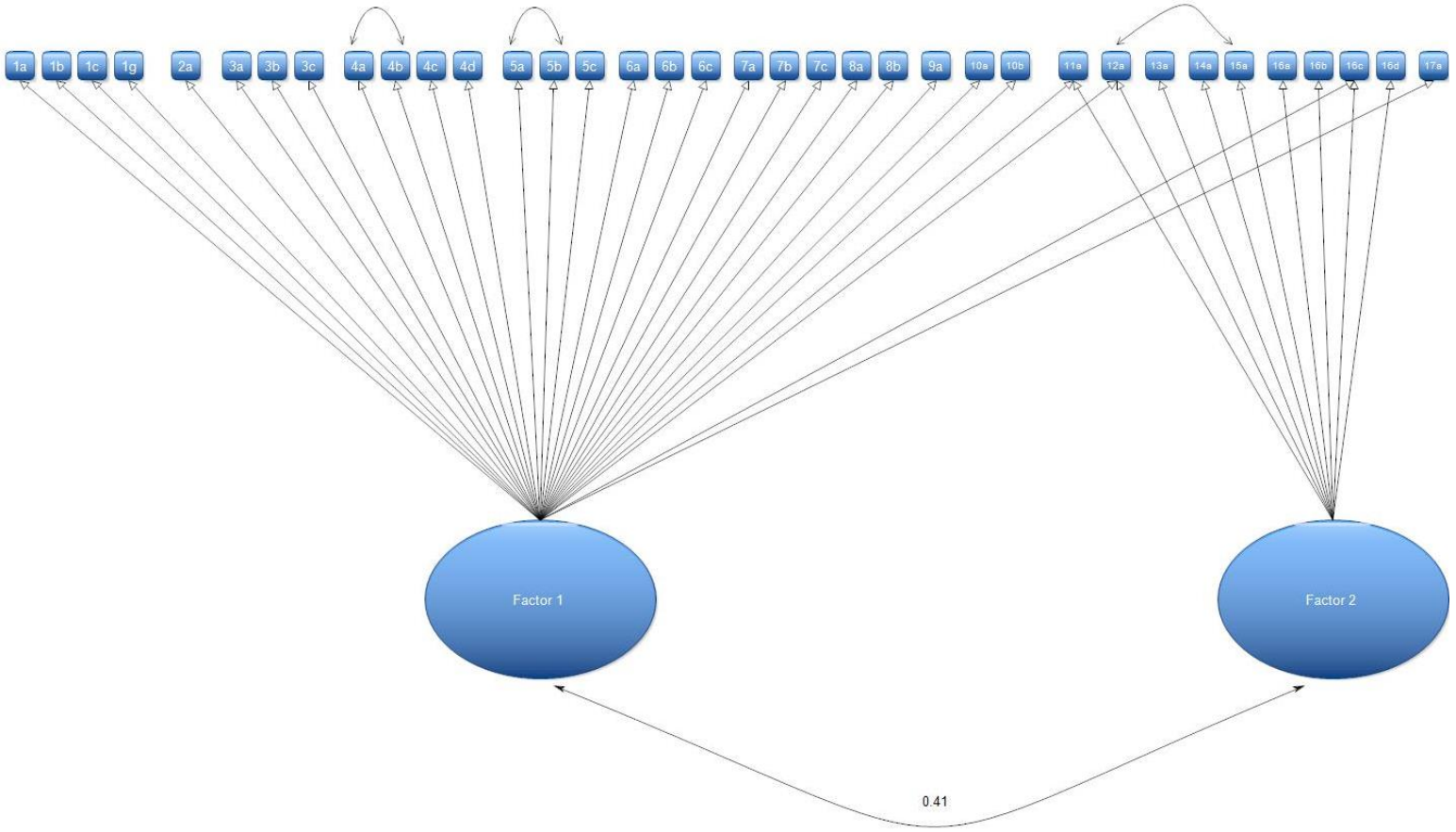


Figure 2. A confirmatory factor analytic model for a two factor model in form R3 items. Items starting with the same number are related to the same stem. The fit indices for this model were: CFI=0.90 and TLI=0.90.

The test information curve for the second factor is depicted in Figure 3. This indicates that, as with the first factor, the information on candidates is maximal below the average level of ability. This suggests the test is more likely to be able to discriminate less able candidates from average to above average ones, but is less likely to be able to discriminate average from above average candidates. The trait relating to factor 1 is approximately normally distributed in the candidate population, whilst that relating to factor 2 shows some evidence of negative skew.

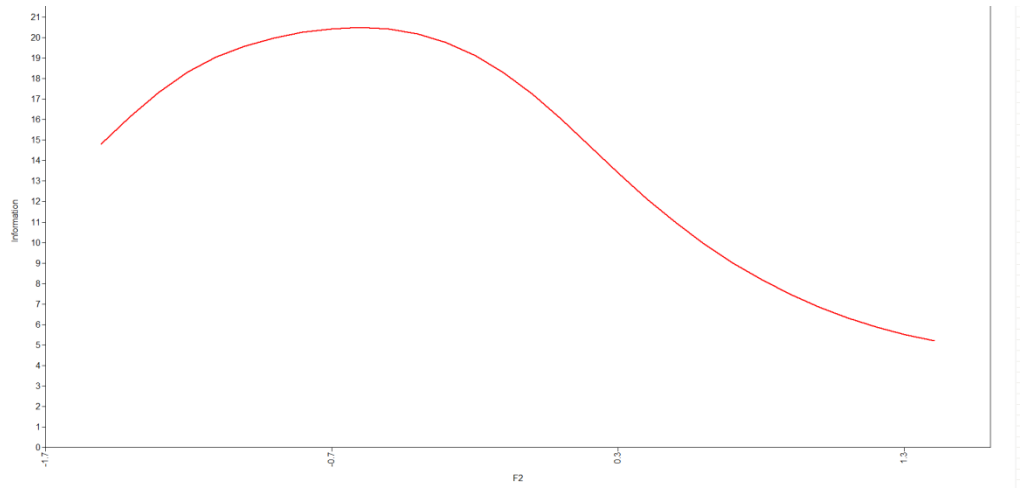


Figure 3. Test information curve (TIC) for second factor from form R3 items. This was derived from a two factor CFA. In order to generate this curve, the CFA was reparameterized into a two parameter logistic (2-PL) model, utilising the graded response model. Thus the curve takes into account both the item difficulties and discrimination values.

Form	McDonald's Omega	Alpha
R1 (N=18 items)	0.73	0.72
R2 (N=19 items)	0.67	0.67
R3 (N=19 items)	0.66	0.67
R4 (N=19 items)	0.73	0.72
R5 (N=19 items)	0.76	0.76
R6 (N=19 items)	0.70	0.69

Table 1. Reliability metrics for the items in the six forms. Note only the items randomly sampled from each stem were included to avoid over-estimating reliability due to item dependency.

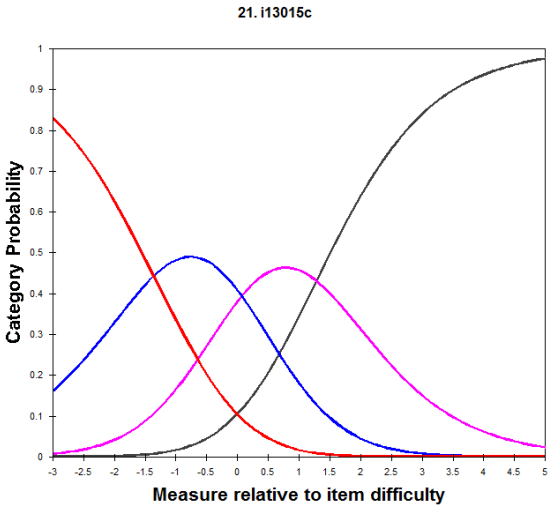
2. Some findings from a Rasch Analysis

Items from form R1 that loaded relatively substantially on the main factor were subjected to a Rasch calibration in order to investigate the extent they could be considered to conform to 'fundamental measurement' principles (i.e. the extent to which responses followed a Guttman pattern of increasing scores with increasing ability- i.e….112122332344344…). The items generally showed a relatively good fit to the Rasch model. Both partial credit models and rating scale models were explored. Partial credit models tend to fit data better than rating scale models, as they have more parameters, although were intended for response data where several steps, each gaining some

credit, were involved in solving a problem (e.g. a maths problem). It should be noted that these assumptions underlying the partial credit model have since been questioned. The person separation indices were similar for both models, and would be considered acceptable as exceeding two, indicating a relatively reliable ability to discriminate between these two groups of testees. It should also be noted that the correlation coefficient between the Rasch calibrated score, derived from items that loaded heavily on the main factor, and the original scores (weighted as per the WPG scoring system) was 0.8655. This infers that around 75% of the variance between the Rasch measure and the original raw scores is shared. Thus, most of the variance in the original, weighted, scores is accounted for by the level of trait or ability reflected by the main factor (most closely to the domain labelled 'integrity'). It is uncertain whether the remaining variance is best conceptualised as error variance or whether it is capturing useful attributes (albeit in an unpredictable manner, in the absence of a well described measurement model). Planned work on predictive (criterion) validity should be able to support or refute this.

In order to investigate the appropriateness of the scoring system currently used, the thresholds were provisionally inspected within a partial credit model. These thresholds represent the ability (trait) level at which a candidate has an equal probability of being observed as responding in either one of two contiguous response categories (e.g. '2' or '3'). Relatively few mis-ordered thresholds were observed (i.e. where candidates generally would be expected to score '3' that scored '2' etc). For example this was observed for 17 of the 62 items in form R1. Those that were observed had thresholds that were quite close together, therefore the mis-ordering may have been within the bounds of the error of measurement. More commonly observed was Rasch-Andrich threshold suppression or 'misordering' (see Figure 8). In such cases, which clearly affected around two-thirds (40/62) of items in R1, intermediate scoring categories were rarely observed. Therefore candidates did not always have the probability of scoring in an expected way, given the estimated ability level. For example the candidate might go from having a high probability of scoring a '1', if the trait was low, straight to a '3' as the intermediate category of '2' was rarely observed (see Figure 2). This makes the intermediate items scores redundant.

a)



b)

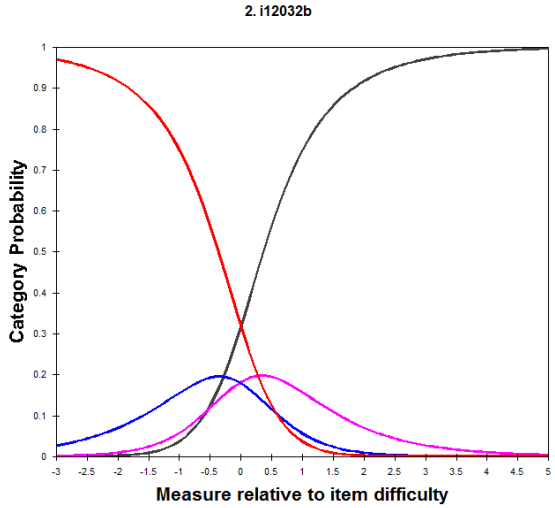


Figure 4. An example of SJT items with well ordered response categories (a) and an example of category 'misordering' (though may not be considered true misordering) due to Rasch-Andrich threshold suppression (b). In the latter case intermediate response categories are rarely observed, meaning candidates of increasing ability tend to transition straight from category '1' to '4'. This implies that no additional information is gained via a four point scoring system in this latter case.

3. Exploring the validity of the weighted scoring system

Using 33 items that discriminated candidates fairly well on the main dimension, the validity of the weighted scoring system was explored. Scoring systems 'A' and 'D' score responses 0, 1, 3, 4 in increasing order (weighted scoring system) whilst systems 'B' and 'C' progress 0, 1, 2, 3 (unweighted scoring system). If the weighting system is plausible then the difference in the average ability levels for those tending to select the second and third most appropriate should be observed as further apart for the weighted scoring system compared to unweighted. By using a Rasch calibration applying the partial credit model we could estimate the average abilities of candidates scoring in each category. Contrary to expectations, the average gap in ability of those scoring in the 2nd versus 3rd response category was significantly wider for those responding to the five unweighted items in form R1 (average gap 0.54 logits, standard deviation (SD) = 0.09) compared to the 28 weighted items (average gap 0.24 logits, SD = 0.07, p for difference <.001). Thus, the practise of weighting the scores of certain items was not supported by this preliminary analysis. Further evidence that would support the weighting of the scores relating to the top two best responses might be if candidates had to be more able to achieve these. On analysis the average ability level of candidates selecting the 3rd category of response (i.e. the 2nd best) for unweighted items was 1.2 logits whereas it was, on average, actually less for weighted items (0.80 logits). This difference was significant at the p<.001 level. There was no difference in the average ability levels of those selecting the best response categories for either weighted or unweighted items (1.04 logits for unweighted versus 1.11 for weighted).

4. Exploring the content of the SJT items in relation to the factor loadings

To better understand how the content of the SJTs related to the dimensionality of the test responses, the authors were granted a viewing of selected items at Pearson Vue offices in Manchester. The content of items in form R1 of the SJT UKCAT test that (a) loaded substantially on factor 1 (a standardised loading of magnitude approximately 0.3 or more), (b) did not load substantially on factor 1, and (c) that loaded substantially on a possible second factor were viewed and a brief thematic analysis conducted.

Table 2 shows the breakdown of 'domain allocation' (i.e. the theme that WPG considered the item to relate to) and response format for the three sets of items viewed. Response format is either *importance* (rating the relative importance of a given factor when considering a social scenario) or *appropriateness* (rating the relative appropriateness of a given social resonse). Not all item domain allocations were available at the time the report was requested, although the majority were provided.

Item loading type	**Response format**		**Original item domain label**			
	Importance	*Appropriateness*	Integrity	Perspective taking	Team involvement	Missing
Loading **heavily on Factor 1** (main dimension) (N=32)	N=29 (91%)	N=3 (9%)	N=15 (54%)	N=8 (29%)	N=5 (18%)	N=4 (13%)
Not loading substantially on Factor 1 (N=23)	N=2 (9%)	N=21 (91%)	N=12 (57%)	N=4 (19%)	N=5 (24%)	N=2 (9%)
Loading on a potential **second factor** (N=11)	N=11 (100%)	N=0 (0%)	N=5 (50%)	N=3 (30%)	N=2 (20%)	N=1 (9%)

Table 2. The item domain labels and response formats for the the three different classes of items, according to their loading patterns

From Table 2 it can be seen that the proportion of items allocated by WPG to each of the three themed domains are very similar for all three classes of items. For example, around half of each set of items are allocated to 'integrity', whether or not the items load substantially on factor 1 or 2. Indeed the most striking difference is the proportion of each response format: almost all of the items loading substantially on factor 1 utilise the *importance* style of response format. The opposite is true for those items that fail to load substantially on factor 1 or load on a putative second factor- these items almost exclusively use the *appropriateness* format. This raises concerns that the factor loadings may be due to a *method effect* (i.e. not relating to a construct being measured, but rather an artefact of the measurement process itself, such as response format or item wording style [e.g. positive versus negatively worded statements]). However, whilst the response format is highly likley to play a role in shaping the structure of the candidate responses, it may be that the response format itself is allowing a different, and some extent orthogonal (i.e. uncorrelated) trait to the main trait to be estimated. On reviewing the content it was evident that the themes of *perspective taking*, *team involvement* and *integrity* could not be easily demarcated (this was highlighted when initially many of the domain labels were missing and we were attempting to guess which theme the item was allocated to). For example, *integrity* was a theme that ran through almost all the items- many

scenarios were based on situations where the candidate (in role) was invited to compromise or maintain their own integrity in different ways. Items allocated to team involvement often included opportunities to compromise integrity, and also demanded a certain amount of ability at perspective taking in order to score highly.

Despite the temptation to dismiss the dimensionality of the SJT UKCAT responses as an artefact of the response method used it should be considered that it is the very difference in response formats that could allow for different traits to be evaluated. Those, mainly *importance*-style, items loading well on factor 1 (the main dimension evaluated by the current SJTs) appeared to test whether the candidate knew why a situation may be socially problematic. In contrast the *appropriateness* items, more related to a putative second dimension (actually most apparent in form R3) appeared to test whether a candidate knew how to respond to such problematic situations, in the most appropriate way.

In theory, if the same situations could be presented to different candiates with different response formats the extent to which dimensionality was attributed to response format rather than content could be calculated (for example, using a multi-method multitrait approach). However, in this present case, as the construct may dictate the response format this may not be possible. It is also interesting to note that those items loading substantially on to a second factor generally involved questions of when, if and how to involve third parties in a situation (e.g. tutors, etc). Thus, they seemed to require some degree of subtlety to making a social judgment beyond a moralistic 'right and wrong'.

5. Discussion

Development of a measurement model, considered robust according to IRT-based metrics, at present, is feasible only if the SJTs that loaded substantially on the main factor are included. This subset of items tend to fit a Rasch model and the resulting scale shows an acceptable person separation index (suggesting that the scale can accurately discriminate between at least two groups of testees). However, the cost of such restriction may be potentially losing almost half the items from the scoring algorithm, and there would be concerns about restricting the item content. Restricting item content may also lead to an increased risk of coaching or practice effects.

It may be possible to mitigate against such coaching or practice effects if both items that do not load substantially on the main factor are also included in the test but are not fed into a scoring algorithm, perhaps even treated qualitatively to facilitate exploration at face-to-face interview. The advantage of such an approach would be to reduce the risk that potential candidates would be able to train for a restricted number of scenarios, as they would not be aware of which items were included and scored and which items were not scored. The disadvantage would obviously be that test time would be taken up by items that would not otherwise add value to the admissions process.

It may be possible to expand to a two-factor model if the content of the item bank is increased. Our brief thematic analysis suggested that there may be a second dimension present in the responses, relating to items utilising an appropriateness response format. These items often relate to situations which may require a judgement about if, when and how to involve third parties in situations. There is the exciting possibility of creating an SJT bank based on a bifactor model representing the dimensions of the ability to know why a situation may be socially problematic and how you go about

addressing. The authors of the report appreciate that generating novel content for SJTs is extremely challenging. The development of a robust measurement model based on two dimensions for situational judgement will provide the advantages outlined above. It will also make it transparent to stakeholders which constructs are being tested, and the measurement precision associate with each of these.

The findings do not appear to generally support the approach of weighting scores for a portion of the items. Some of the items appear to be suitable for having four separate scoring categories, while many do not. However, moving to a purely dichotomous scoring system is likely to result in informational loss and is probably best avoided.

The information yielded by the test is maximal at the lower range of ability. This supports the use of the SJT scores as a 'screen-out' test within the selection process. It is plausible to offer more detailed breakdown of scores in the future to universities, if that is what they are requesting, on the understanding that is unlikely that scores will be able to accurately discriminate between candidates at the average to above-average range of ability. For example, if deciles for scores are provided then it would be defensible to 'screen out' those candidates in the lower two to three deciles, but it would be less certain whether those in the higher deciles could be reliably dicriminated from one another. The addition of more difficult items could shift the test information curve to the right (i.e. deliver higher levels of information on more able candidates).

Whilst UKCAT consider that the potential practise of feeding back only two bands of scores to universites may be risky, there is little rationale or evidence to support the reporting of scores by breaking them into four bands - especially as there is likely to be insufficient information to discriminate candidates in the the top two or three bands accurately. Alternatives include reporting back the scores broken into deciles (with a 'health warning' that the test is unlikely to be reliable above the fourth decile or so). A second alternative would be to report back the scores as standardised z or T scores with the advice that the test should not be expected to reliably differentiate candiates above the mean score (i.e. z score of 0 or T score of 50). This would have the advantage of identifiying very low scoring candidates easily (e.g. z score of less than -2.0).

Acknowledgements

Thanks to Harriet Rimington from Pearson Vue for facilitating access to the data and item content and to the UKCAT Board for funding the analysis.