Report to the UKCAT board: enhancing the scoring of the situation judgement test component of the UKCAT; reanalysis of the validity pilot study data

Summary report

Paul Tiffin and Lewis Paton, University of York, October 2016

(Summary report produced March 2017)

Executive summary

Background

Previous work analysing item level data from the situation judgement tests (SJTs) suggest that reliability could be potentially increased by using an unweighted scoring system, selecting a subset of items that relate strongly to the main dimensional being measured and exploring scoring patterns using item response theory (IRT) approaches. However, it was unclear whether increasing the reliability of the test would automatically increase validity, as it may have been possible that important test content could be lost taking this approach. The original validity pilot study data provided an opportunity to evaluate competing scoring approaches in respects of both reliability and validity.

Methods

The relationship between the tutor ratings (the validity criterion) and the original equated SJTs scores varied substantially across each form of the SJTs used in 2013. Therefore the test forms were analysed separately. A stepwise approach was taken in order to ascertain the effects of changing the scoring system. Firstly an unweighted scoring system was used (0, 1, 2, and 3 for each response category). Secondly, the items pertaining to each form of the test were subjected to a Rasch analysis using a partial credit model (PCM). Thirdly items with disordered scoring categories (according to the PCM) were recoded. The items were then re-analysed using a Rasch model. Fourthly, items loading heavily (the magnitude of approximately 0.3 or more) onto the main dimension being measured by each form were selected and retained before, again, being subjected to a Rasch analysis. Finally, the test forms were subjected to analysis taking a multidimensional item response theory (MIRT) approach where five main traits were postulated as being related to the item response patterns.

Results

There was substantial variation in validity across the test forms. Using unweighted scores improved the validity coefficients in two forms of the test, degraded it in approximately two forms and made no substantial difference in the remaining two forms. The items for each form were Rasch calibrated and all the constituent items showed acceptable fit to the Rasch model. The Rasch calibrated scores for the participants generally increased the validity coefficients in between two and four of the test forms (depending on the outcome- *integrity*, *team involvement* or *perspective taking*). Recoding of the apparently misordered score categories further improved the validity resulting in four out of the six forms having validity coefficients exceeding those of the original equated SJTs scores. Selecting only items that loaded on the main dimension being measured dramatically improved the validity coefficients for form two of the test, somewhat improved validity in form one, but degraded it in the remaining four forms of the test.

Factor scores derived from the multidimensional item response modelling were able to predict more of the variance in the tutor ratings than the original SJTs scores. This approach aids understanding of the test characteristics but is unlikely to be useful for designing *a priori* scoring systems. There was little evidence that Extreme Response Style (ERS) substantially influenced the performance of candidates on the test.

Conclusions

The issue of equating the SJT scores must be addressed as a matter of priority. Subjecting items to Rasch calibrations by form, in order to revise scoring in some cases, *post hoc*, generally improves the validity of the resulting scores. Retaining items that relate to the main dimension/s being measured has the potential to dramatically improve the validity of SJTs but, importantly, only in cases where the main construct being measured by the test or test form is substantially related to the construct that it is been validated against. The results of these analyses highlights the challenges of performing effective equating with SJTs in the absence of well described measurement models. Several potential approaches for addressing this issue are described.

Background

Previous work analysing item level data from the situation judgement tests (SJTs) suggest that reliability could be potentially increased by using an unweighted scoring system, selecting a subset of items that relate strongly to the main dimensional being measured and exploring scoring patterns using item response theory (IRT) approaches (Tiffin & Carter, 2013). However, it was unclear whether increasing the 'reliability' (i.e. increasing the information yielded on each candidate) of the test would automatically increase validity, as it may have been possible that important content could be lost taking this approach. The data relating to the original validity pilot study conducted by Work Psychology Group (WPG) provided an opportunity to evaluate competing scoring approaches, and the extent to which they increased validity of the test as well as reliability (Patterson, Edwards, Rosselli, & Cousins, 2015).

Methods

The details of how the original data were collected are outlined in the original report by WPG to the UKCAT board (Patterson et al., 2015). However it is important to mention that four medical schools participated in the pilot study and the main outcomes were related to tutor ratings. These ratings were provided in the form of a percentile rank for each student related to a tutor. Tutors were asked to rate their students across three domains: *perspective taking* (PT), *team involvement* (TI) and *integrity* (IN). A mean tutor rating was also recorded though is not analysed here. As the ratings tended to correlate to a high degree ($r>0.72$) then additional information is unlikely to be acquired by additional analyses with the mean tutor rating as an outcome. Tutors were also asked to produce a rating of whether students were *particularly promising*, *average* or *likely to struggle*.

The supervisor ratings (outcome variables) were roughly normally distributed as can be seen, as an example, for integrity ratings in Figure 1 with the accompanying quantile plots ('Q-Q' plots). For this reason parametric methods were used to evaluate the relationship between the SJTs scores and the supervisor ratings. However, due to the small numbers of students having tutor ratings and having taken each of the six forms of the SJTs, standard errors were derived via bootstrapping (with 1000 replications) to accommodate the relatively small number of observations in each group. The original equated SJT scores were regressed on to the tutor ratings. The results are depicted in Table 1, split by form.

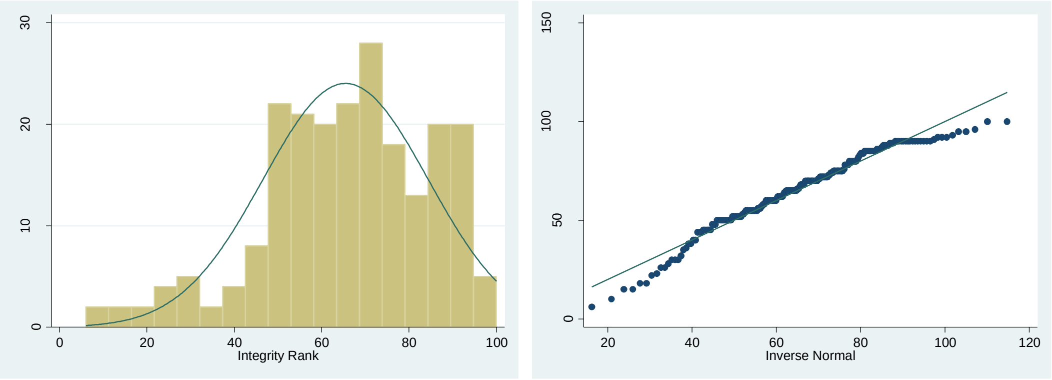

Figure 1. Distribution of supervisor ratings for *integrity* with accompanying quantile plot (normal distribution).

Test form	Standardised coefficient	P value	Lower 95% CI	Upper 95% CI
		Integrity		
1 (n=45)	0.35	0.03	0.04	0.65
2 (n=30)	0.34	0.03	0.04	0.65
3 (n=43)	0.22	0.11	-0.05	0.48
4 (n=30)	0.08	0.63	-0.25	0.41
5 (n=34)	0.29	0.02	0.05	0.52
6 (n=36)	0.06	0.71	-0.28	0.40
		Team Involvement		
1 (n=45)	0.34	0.02	0.05	0.62
2 (n=30)	0.22	0.22	-0.13	0.57
3 (n=43)	0.22	0.06	-0.01	0.45
4 (n=30)	0.14	0.45	-0.23	0.51
5 (n=34)	0.09	0.60	-0.24	0.42
6 (n=35)	0.23	0.09	-0.04	0.49
		Perspective Taking		
1 (n=45)	0.34	0.02	0.06	0.62
2 (n=30)	0.27	0.10	-0.05	0.60
3 (n=43)	0.13	0.31	-0.12	0.39
4 (n=30)	0.11	0.53	-0.24	0.46
5 (n=34)	0.10	0.53	-0.21	0.40
6 (n=36)	0.07	0.66	-0.26	0.41

Table 1. Regression coefficients for the original SJT scores, split by form.

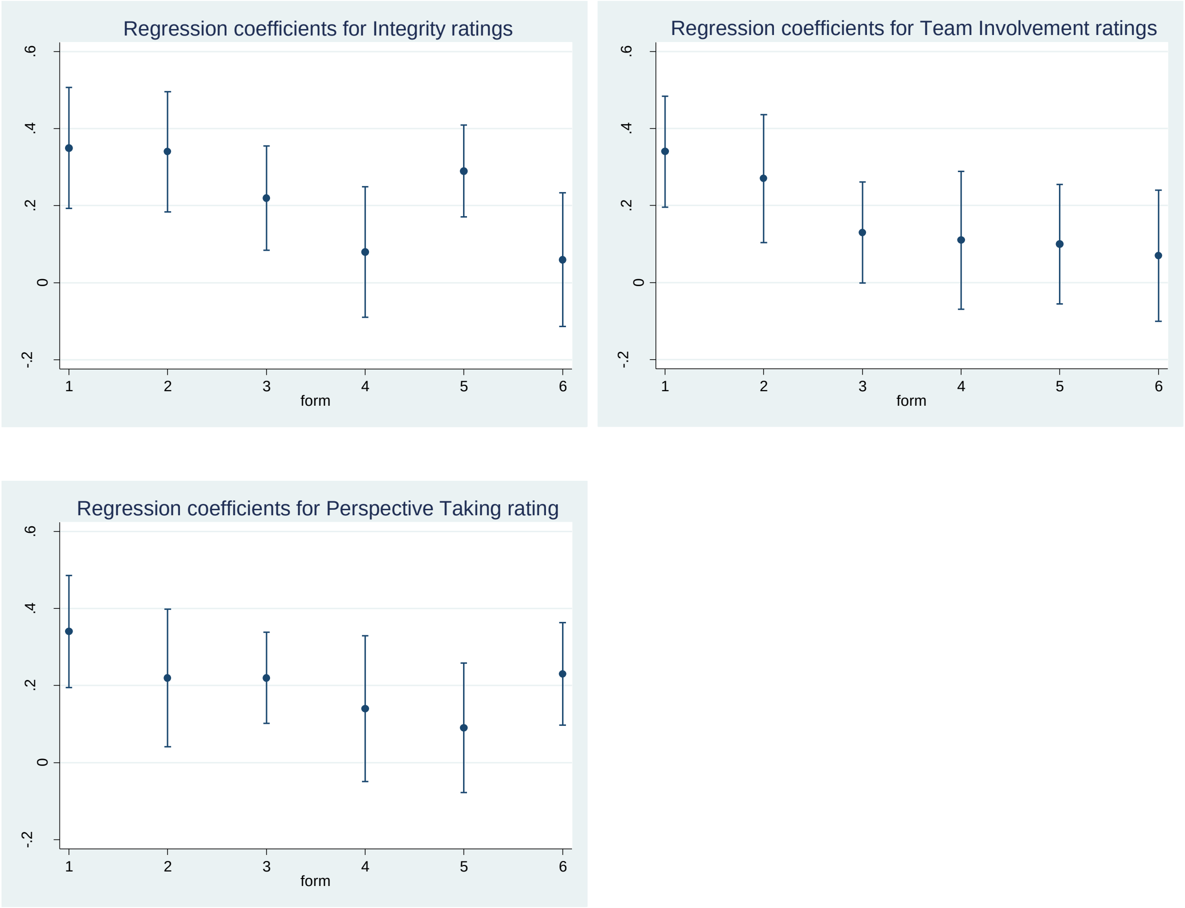

Figure 2. Regression coefficients and standard errors for the tutor ratings predicted from SJT scores for each of the six forms of the test.

From Table 1 and Figure 2 it can be seen that there is marked variation in the validity coefficients between forms. A one way analysis of variance to evaluate differences in the coefficients across forms returns a p value of 0.08. Whilst this is not statistically significant at the p<.05 level- there are few observations (n=18- six forms and three ratings) to base this test on- we can be reasonably (92%) certain that the variation in these observations is unlikely to be due to chance alone. Pairwise comparisons also highlight particular differences between the coefficients between specific forms (e.g. form two and form six). It can be observed from Figure 2 that the variation of the coefficients across forms appears more marked than across outcomes (IN, TI and PT) within forms. Indeed the coefficients are not significantly associated with the rating type (IN, TI and PT). This is unsurprising as the three main rankings themselves are relatively highly correlated (Spearman's rho values range from 0.72 to 0.81). The impact of this lack of equivalence between forms can be further illustrated by the following example. The mean SJT scores for those with an *Integrity* ranking of less than 50 was about 201 (n=4) and 207 for those above 50 (n=27). However, for those taking form six with low rankings (<50) on the *integrity* ratings the average equated SJT score was 219 (n=4) and those with higher ratings actually lower at 213 (n=33). Thus there were marked differences between the SJT scores on 'high' and 'low' tutor ratings depending on the form of the test taken. In light of these findings it was decided to analyse the six forms of the SJTs separately. For convenience and clarity, hereafter in this report, the equated SJTs scores will be referred to as the original SJT scores.

The original scoring systems included weighting, according to an expert opinion, for certain response categories (e.g. a scoring sequence 4, 3, 1, 0 for ordered responses). In order to begin to analyse the scoring system we did not assume that particular responses should receive extra weighting and therefore explored the item responses using an unweighted scoring system (e.g. a scoring sequence 3, 2, 1, 0). Likewise, the occasional items that were scored originally using a 'tied' system (e.g. 3, 2, 2, 0) were untied to facilitate exploration using item response modelling.

For the item response modelling, each set of items relating to each of the six forms were subjected to a Rasch analysis in the WINSTEPS software package using a partial credit model (PCM). The PCM allowed the identification of items where there were suggestions of misordering (i.e. those with higher estimated abilities tended to score lower points, on average). Whilst the assumption of unidimensionality was unlikely to be well supported within each set of response patterns in each form this does not necessarily impact on the ability of an item to fit the Rasch model. Indeed, an item's fit to a Rasch model is determined by the Guttman sequence exhibited in the responses (i.e. a progression in this case from 0's to 3's as estimated candidate ability (theta) increases). Moreover, it may be that multidimensionality is less likely to impact on the fit of items in a test to the Rasch model where the dimensions are probably correlated to some degree. Items showing evidence of misordering were recoded. A second round of Rasch analysis and recoding was conducted in order to identify any remaining items with evidence of substantial misordering (except in the case of form 4, where only one round was required to eliminate any evidence of misordering). These re-scored items were then subjected to a further Rasch analysis.

In order to ascertain the main dimensions underlying the item responses each form was subjected to a series of ordinal exploratory factor analyses which were conducted in Mplus v7.4. These were performed in combination with a geomin rotation and used full information maximum likelihood ('direct ML') as the estimation method. Only those items with a standardised rotated loading of approximately three (≥0.27 in this case) were included in a final Rasch analysis.

Multidimensional item response theory is being increasingly applied to situational judgement test response data, due to the rather ‘fractal’ multidimensional nature of SJT scores. For this reason multidimensional item response theory-based (MIRT) modelling was carried out on the response data. Exploratory factor analyses were conducted in Mplus using weighted least squares (robust to variations in mean and variance) as the estimation method, exploring a postulated five factor structure in each form. Note that our previous work suggests that the SJTs using the UKCAT generally have one, or at most two main dimensions underlying their response structure. However, in this case we included five postulated factors as they provided a better fit to the data and we also wanted to explore the extent to which more minor factors, or ‘artefactors’ (e.g. ‘minifactors’ related to dependency within items or similar wording), may be related to the validity criteria. Due to the “fuzzy multi-dimensional” nature of the SJT response data adequate fit (i.e. Tucker-Lewis Index>0.90), even on exploratory ordinal factor analysis, was generally only achieved by the inclusion of 10 or 12 factors. However such complex structures were not amenable to the usual estimation methods and therefore the number five was selected as the number of factors to be retained as a compromise, given the computational resources available. Technically, the MIRT models were implemented in Mplus using full information maximum likelihood as the estimation method. Due to the complexity of the models Monte Carlo integration was used with 50 integration points. This was increased to 100 integration points for models where estimation failed to converge using 50 integration points. Items were allowed to load onto one of the five factors in each MIRT model relating to the six forms, if they have been found to have a rotated loading of magnitude greater than 0.3 on the prior exploratory ordinal factor analysis.

Both univariable and multivariable analyses were conducted exploring how each of the five traits might be related to the tutor ratings in each of the six forms. We also compared the amount of variance in the tutor ratings that can be accounted for by these traits combined compared to the original SJT scores. Such exploration was not intended to guide future scoring approaches as such, but was intended to feed into test development by highlighting which factors (and thus related items) were most closely related to the validity criteria.

Results

The results comparing the predictive validity, by form, of the various scoring methods of the 2013 UKCAT SJTs are summarised in Table 2.

Integrity					
Form	Original SJT scores	Unweighted sums	Rasch scores	Rasch with rescoring*	Rasch- selected items only
1	0.35	0.31	0.32	0.31	0.30
2	0.34	0.34	0.36	0.39	0.58
3	0.22	0.24	0.24	0.30	0.06
4	0.08	0.10	0.10	0.10	0.01
5	0.29	0.28	0.28	0.35	0.14
6	0.06	0.05	0.03	0.04	-0.05
Team involvement					
1	0.34	0.29	0.32	0.30	0.40
2	0.22	0.23	0.22	0.24	0.35
3	0.22	0.24	0.24	0.35	0.16
4	0.14	0.13	0.15	0.15	-0.05
5	0.09	0.09	0.09	0.09	-0.04
6	0.23	0.17	0.17	0.15	0.17
Perspective taking					
1	0.34	0.31	0.33	0.31	0.35
2	0.27	0.28	0.30	0.32	0.43
3	0.13	0.16	0.15	0.24	0.08
4	0.11	0.11	0.11	0.11	0.03
5	0.10	0.09	0.09	0.14	0.05
6	0.07	0.03	0.01	-0.01	-0.02

Table 2. A summary of the validity coefficients for the six forms of the SJT and the various rescoring methods. Those values highlighted in green exceed those from the original scoring system (note that values are rounded to two decimal points).

*That is rescoring of apparently ‘misordered’ items on Rasch analysis

It was observed that the Rasch-based scoring system (with some rescoring of misordered items) performed better when each form was analysed separately. Indeed, the Rasch scores for the selected items for form two showed a very high correlation with integrity ratings, approaching 0.6. However, when the whole population (irrespective of form taken) was analysed the original scoring tended to be slightly more predictive (Table 3). This somewhat puzzling, paradoxical observation can be potentially explained as follows; the Rasch based scoring is tied to the ‘Rasch dimension’ that each form exhibits (i.e. as candidates get more able they will tend to score more highly). Items that don’t follow this pattern can be observed as showing misordered scoring, or misfit to the Rasch model. Scoring can be recoded to correct for this. However, this Rasch dimension appears to vary across the different forms of the test. Thus, on a form-specific basis the relationship between the scores and a specific outcome (e.g. *integrity* rating) can be increased by this approach in most cases. However, because the forms are not equated in the true sense (i.e. equal scores equate to equal ability levels across forms) looking at the relationships with the validity criteria across all forms at once will tend to reduce the overall validity coefficients. Thus, although individually the signal to noise ratio is increased for each specific form, the underlying construct that the signal is related to is different in each form, paradoxically leading to reduced overall validity of pooled scores. However, a Rasch measure could be constructed by analysing the entire dataset simultaneously (items not included in forms taken by candidates were treated as missing data), with all items showing acceptable fit to the Rasch model and a person separation index (a Rasch index of reliability) approaching 2.0. Again, two iterations of correcting apparently misordered items scores were conducted. This resulted in a scale that outperformed the original SJT scores in predicting the tutor ratings (see Table 3). In addition this overall Rasch calibration with correction for apparent misordered codings performed as approximately as well as the original SJT scores in predicting the overall banding of the candidate on an ordinal logistic regression analysis (Table 4). We also noticed, on graphing the rating data, that there was a ‘tail’ of students with rankings of lower than 40. We therefore used logistic regression to model the relationship between ‘low’ ratings (by this definition) and the three ‘test-level’ (as opposed to ‘form-level’) different test scoring approaches (Table 5).

Multidimensional Item Response Modelling

In multidimensional item response theory it is postulated that more than one trait can be related to the candidate’s responses to test items. Classically, in confirmatory factor analysis a simple structure is sought, where one item typically loads heavily on one factor, and close to zero on any others. This is not the case in multidimensional item response theory where items are allowed to cross load. Thus, for example, a candidate may need to have high trait levels on several dimensions in order to achieve high scores on certain items.

Integrity				
Scoring method	Std. coefficient	P value	Lower 95% CI	Upper 95% CI
Original	0.25	<0.001	0.14	0.37
Rasch-form specific	0.23	<0.001	0.12	0.33
Rasch- whole test	0.28	<0.001	0.17	0.39
Team Involvement				
Original	0.22	<0.001	0.09	0.34
Rasch-form specific	0.20	<0.001	0.08	0.32
Rasch- whole test	0.22	<0.001	0.10	0.35
Perspective Taking				
Original	0.20	<0.001	0.08	0.32
Rasch-form specific	0.17	0.003	0.06	0.29
Rasch- whole test	0.22	<0.001	0.11	0.34

Table 3. Results across all forms, prediction of tutor ratings, of the UKCAT 2013 SJTs, comparing the original SJT scores, the form specific Rasch scores (adjusted for misordered scores) and a Rasch measure constructed from the entire test.

Overall rating (1=promising, 2=average, 3=likely to struggle)				
Scoring method	OR	P value	Lower 95% CI	Upper 95% CI
Original	0.98	0.02	0.96	1.00
Rasch-form specific	0.61	0.18	0.29	1.26
Rasch- whole test	0.38	0.02	0.17	0.86

Table 4. Results of an ordinal logistic regression analysis across all forms for the prediction of ‘overall’ rating of students by UKCAT 2013 SJT scores, comparing the original SJT scores, the form specific Rasch scores (adjusted for misordered scores) and a Rasch measure constructed from the entire test.

Scoring method	Coefficient	P	95% lower CI	95% upper CI
Low Integrity				
SJT-Original	0.95	0.002	0.92	0.98
Rasch-form sp.	0.14	0.01	0.03	0.57
Rasch- all	0.07	0.001	0.01	0.35
Low Perspective Taking				
SJT-Original	0.98	0.29	0.95	1.01
Rasch-form sp.	0.52	0.22	0.18	1.50
Rasch- all	0.43	0.24	0.11	1.73
Low Team Involvement				
SJT-Original	0.98	0.06	0.95	>1.00
Rasch-form sp.	0.36	0.05	0.13	0.99
Rasch- all	0.32	0.09	0.09	1.18

Table 5. Results from a logistic regression predicting *low* category of tutor ratings (percentile ranked less than 40 for IN, PT and TI) across all forms. Original scores are compared against form-specific Rasch scores (‘form sp.’) and a whole test-based Rasch score (‘all’).

Integrity ratings- R^2 values			*Team involvement* ratings- R^2 values		
Form	SJT	MIRT	Form	SJT	MIRT
1	0.12	0.15	1	0.11	0.24
2	0.12	0.39	2	0.05	0.21
3	0.05	0.18	3	0.05	0.24
4	0.007	0.07	4	0.02	0.24
5	.09	0.29	5	0.008	0.26
6	.004	0.11	6	0.05	0.08
Perspective taking ratings- R^2 values			Overall ratings- R^2 values		
Form	SJT	MIRT	Form	SJT	MIRT
1	0.12	0.19	1	0.12	0.19
2	0.07	0.28	2	0.07	0.28
3	0.02	0.17	3	0.02	0.17
4	0.01	0.04	4	0.01	0.04
5	0.009	0.24	5	0.009	0.24
6	0.006	0.14	6	0.006	0.14

Table 6. The amount of variance in each set of rating explained by original SJT scores and by the traits estimated via the MIRT models for each of the six forms, as indexed by the R^2 values.

The results from univariable regression analyses, including with bootstrapped standard errors to accommodate the small number of observations (not included in this report) were provided to WPG in order to feed into test development. It was hoped that information on how the different factors in each form (and their related items) relate to the tutor ratings would help guide future item selection and content creation. It can be seen in Table 6 that the five traits recovered from the MIRT models consistently explain more of the variance in the tutor ratings than the original SJTs scores. The multidimensional item response modelling can thus be thought of as a process by which some of the noise is removed from the raw scores and the signal from the relevant traits or abilities is allowed to surface.

Extreme Response Style and UKCAT SJT scores

It is well known that responses to questionnaires and tests using Likert or Visual Analogue Scales may be influenced by a candidate's response style. This includes an 'extreme response style' (ERS) where extreme responses tend to be preferred over centrally located ones (e.g. *very inappropriate* vs *somewhat inappropriate*). The UKCAT SJT has four point Likert scales as the response format and therefore at least a brief exploration of the potential of ERS to influence scores should be considered. For this purpose we took form two, which has a main dimension strongly linked to the validity construct. Emerging research suggests that there are a number of ways the potential of ERS to impact on scores can be investigated. The most practical and perhaps most effective option is currently to create indicators of ERS which 'shadow' the actual items (i.e. a 0 or 3 response is coded '1' while a 1 or 2 is recoded as a '0'). These 'shadow indicators' can be jointly modelled in a factor analysis (in effect, a multidimensional IRT model) and the covariance of the 'ERS factor' with the main factor of interest can be evaluated (Wetzel, Böhnke, & Rose, 2015). Where low correlations with ERS are observed this is unlikely to be a substantial problem, and it will be of little practical value to correct scores for this. In the case of form two of the SJT we observed that an ERS factor correlated 0.27 with the main factor measured. This is relatively low. To further investigate this, the ERS factor scores were recovered and put into a multivariable model, along with the main factor scores. After controlling for the influence of the main factor being measured, ERS was observed to have little relationship with tutor ratings (adjusted betas ranged from 0.14 to -0.09). Thus the influence of ERS on the SJT scores is likely to be relatively trivial.

Discussion

The findings from this re-analysis were in keeping with emerging work relating to the modelling of SJTs scores. For example, a recent thesis reported an exploration of taking different approaches, including multidimensional item response modelling, to the scoring of SJTs (Whelpley, 2014). The author concluded that although the application of item response theory shows promise for improving the performance of SJTs no individual scoring method consistently produces the highest levels of validity across all sets of items. In the present case item response modelling was relatively easily applied, as the responses were in Likert scale format. Although, due to the small number of observations, study power was generally lacking, it was clear that there was substantial variation in the validity across forms of the UKCAT 2013 SJT. For example, scores from form 2 (especially when factor scores were recovered via IRT) were strongly related to tutor ratings. In contrast the scores from form six showed virtually no relationship between tutor ratings for *integrity* and *team involvement*. Thus analyses had to be conducted separately for each form of the test. In psychometric terms, test equating implies that equivalent scores on various forms of a test would indicate the same level of trait or ability (theta) in candidates. In the absence of a well-defined and understood measurement model (i.e. where the measurement of theta by a set of indicators (items) is understood) it would be extremely challenging to assure equating to an acceptable level of confidence. Moreover, it may be the situation that the distribution of scores between test forms appears to be identical and even internal reliability consistency, as indexed by Cronbach’s Alpha, may be relatively similar. Nevertheless this does not necessarily imply that the test forms are measuring the same constructs to a comparable degree. Thus, it is possible that we may have a situation where a candidate’s score is mainly determined by the form of the test they are randomised to. At the population-average level, when dealing with large numbers, this will be unlikely to be of importance, as traits can be assumed to be normally distributed across the population. However, at an individual, candidate-level, a candidate who is high on one trait and would have done well on one form of the test may do significantly more poorly on another form. Thus, unless a single test form is used or equating is relatively well assured, test results may be open to challenge by candidates who failed to secure a place at medical school due to their performance at the UK SJT. At present the position of UKCAT (and thus Pearson Vue and Work Psychology Group) is, in our opinion, relatively defensible in that candidates were randomised to various forms of a test, where the scores were re-scaled to allow for the overall difficulty of the items. It was plausible to assume that, despite the ‘fractal multidimensionality’ of the SJT responses the same underlying constructs were generally being tapped into by all forms. It was interesting to note that when all test items were Rasch calibrated simultaneously a relatively good fit was observed, despite the underlying multidimensionality. This did suggest that the traits being tapped into were generally correlated and ‘pushing’ the Guttman sequence in the right direction (i.e. that more generally able candidates tended to score more highly on items). However, these analyses suggest that substantial variation in validity across forms exists, at least in the 2013 SJTs. Thus, this new knowledge brings with it an imperative and responsibility to address this challenge in future testing rounds to increase the effectiveness of the SJTs and ensure fairness, as well as maintain a defensible practice.

In keeping with the previous analysis of SJTs scoring there seems to be no consistent advantage of weighting certain scores. We noted that the validity coefficients tended to degrade for forms one and six when the scores were ‘unweighted’, so it may have been that more difficult items in these forms of the SJTs were the ones that experts decided to allocate additional scores to. What seems to be the case is that were the construct being measured by the test form overlaps with the construct that is estimated by the validity criterion then recovering factor scores reduces the measurement ‘noise’. This results in fairly substantial increases in validity. However in cases where the main dimension being measured by test form does not coincide with the validity construct than this approach actually degrades validity. This concept is illustrated in Figure 3. Whilst this finding may be somewhat frustrating- that one size of scoring method does not fit all forms- it highlights the potential for dramatically improving the validity of SJTs. In particular, the main dimension being measured by form two of the UKCAT 2013 SJT appears to have tapped into the trait that is also estimated by tutor ratings. When this trait was estimated using factor analytic techniques some of the validity coefficients approached 0.6. These kind of close correlations are unusual to find in social sciences research. The results of analyses also provide clear guidance on how to build on item content that is likely to enhance validity. This not only holds the potential to achieve high levels of criterion validity but also may make test equating between, at least two or three, forms of the test feasible.

The findings from the multidimensional item response modelling does not immediately assist with finding an *a priori* scoring system- in the absence of evidence for validity it is uncertain which the five traits models are most strongly related to the tutor ratings and how they might interact. However the findings are likely to help test developers select items that are most valid, as these will be related to the traits that are most strongly predictive of tutor ratings on the multidimensional item response modelling. The results of multivariable regression modelling with these traits show that quite substantial portions (up to about 40%) of the variance tutor ratings can be explained by these. This is all the more surprising when considering some of the challenges in obtaining informative supervisor or tutor ratings of subjective concepts such as ‘integrity’. Indeed, WPG are to be commended on obtaining high quality, and discriminating, tutor rating across a number of medical schools. In future it may also be helpful to include a unique identifying id for tutors so that any dependency (clustering) of ratings within tutors can be adjusted for in statistical analyses.

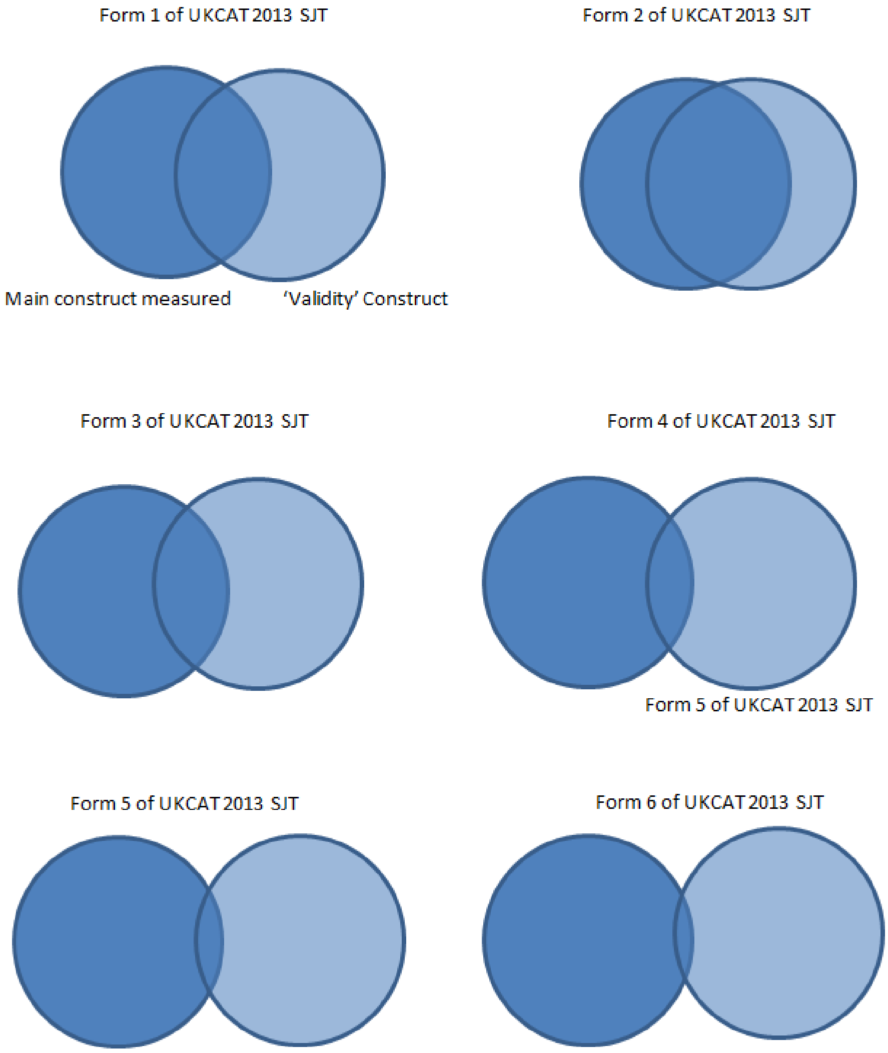

Figure 3. Venn diagrams illustrating how the main dimension being measured by each of the UKCAT 2013 SJT test forms may overlap with the validity construct (note overlaps not strictly to scale). Those with most overlap are likely to have the validity of the scores increased by IRT approaches to scoring.

Hopefully this report will give some guidance about both potential changes to scoring systems and further test development.

Recommendations

1. That the issue of equating the UK SJT be addressed as a matter of priority, given the implications for test effectiveness and fairness. Currently three forms of the test are used. There are several options to address the issue of equivalence:

a. One obvious option is to consider using a single form of the test for each sitting though this option brings certain risks, such as test content being memorised and leaked. This risk is appreciable given the relatively long testing window for the UKCAT SJTs.

b. A second option is to have two or three forms with a considerable overlapping ‘core’ of shared items. Only the shared items would be put into a scoring algorithm. The main problem with this approach would be that precious allocated test time is being used for items that are not being scored. On the other hand, if the items that are known to have good psychometric properties are scored in the ‘core pool’ then relatively high levels of reliability, test information and validity could be achieved with a relatively small number of test items. This option also brings some security risks if candidates are able to recognise and memorise shared items.

c. Given the challenges of test equating it may be considered to reduce the number of forms to two, with considerable overlap between forms. By increasing the content of several test forms with items are that are known to load heavily on the main factors related to the validity criteria, and allowing for the considerable overlap between test forms, it may be possible to achieve test equating between two (or perhaps even three) forms by establishing a robust measurement model for the SJT. This is the most complex and riskiest of the options but is feasible. Attempts at ‘blue printing’ by balancing the proportion of items labelled with specific domains (e.g. ‘integrity’) may improve equivalence to some extent as certain domains may be more associated with certain latent variables (factors) than others. However, it is how items load on to these factors, not allocated to the domains *per se,* that is of crucial importance when working to improve equivalence between test forms.

d. In theory, if a sufficiently robust criterion could be found, then a machine learning approach could be used to predict the later performance from the SJT responses, side-stepping the equating issue. Machine learning (‘artificial intelligence’) approaches are concerned with *prediction* rather than *explanation* (hence sometimes referred to as ‘black box’ methods). However, in this case an algorithm would have to be trained on a set of SJT reponses against a criterion and the qualities of both the outcome and predictors (SJT items) could shift over time, becoming less valid.

2. That an unweighted scoring system be considered for the UKCAT SJTs so that scoring can be checked post hoc via a Rasch partial credit model. This process might highlight cases where misordering of scores has occurred and increase both reliability and validity of the test. Moreover, Rasch calibrated scores tended to modesty outperform the original scores at both form and test level and may be used in preference to raw scores.
3. That the items relating to the factors that have the strongest relationship with the validity criteria be included in future tests. The relevant items should be examined to see if it is possible to build further content which relate to these. In this sense item response modelling should feed into test design in an iterative manner. If more robust measurement models are achieved for the SJT then the application of IRT based models could substantially enhance the validity of the scores.
4. Further work exploring the validity of the 2013 SJTs needs to consider and incorporate these findings. For example, analyses related to the predictive validity of the 2013 SJTs should include initial analyses with a breakdown by each test form- this will provide further evidence relating to the equating of the test forms. It may be that certain forms that have a relatively weak relationship with a specific validity criterion should be excluded from such analyses, or that scores from forms with similar profiles (e.g. forms one and two) could be pooled.
5. Previous analyses relating to predictive validity, where different forms of SJTs have been used in medical selection, should be re-run with analyses reported by test form to see if the potentential equating issue also affects these results. Obviously dividing observations by test form degrades study power and suitable statistical approaches and more cautious interpretation may be required when appraising the subsequent findings.

References

Patterson, F., Edwards, H., Rosselli, A., & Cousins, F. (2015). *UKCAT SJT Predictive Validity Study Summary Report*. Retrieved from Derby, UK: http://www.ukcat.ac.uk/App_Media/uploads/pdf/Understanding%20the%20measurement%20model%20of%20the%20UKCAT%20SJT.pdf

Tiffin, P. A., & Carter, M. (2013). *Understanding the measurement model of the UKCAT Situational Judgment Test: Summary Report*. Retrieved from http://www.ukcat.ac.uk/App_Media/uploads/pdf/Understanding%20the%20measurement%20model%20of%20the%20UKCAT%20SJT.pdf

Wetzel, E., Böhnke, J. R., & Rose, N. (2015). A Simulation Study on Methods of Correcting for the Effects of Extreme Response Style. *Educational and Psychological Measurement*. doi:10.1177/0013164415591848

Whelpley, C. E. (2014). *How to Score Situational Judgment Tests: A Theoretical Approach and Empirical Test.* Virginia Commonwealth University, Richmond, VA.