Work Psychology Group
Thinking differently

UKCAT SJT: a study to explore validation methodology and early findings

July 2014

Prof Fiona Patterson
Stuart Martin

1. Executive summary

1.1 This report describes a study that explored a methodology with which to investigate the predictive validity of the UKCAT SJT.

1.2 The key objectives of this study were to a) identify issues to assist the development of a robust specification for a longitudinal predictive validation study, and b) provide an early indication of the validity of the SJT in the context of two participating medical schools.

1.3 In-training performance data were gathered using a bespoke questionnaire that was issued to GP, GP Community and Problem Based Learning (PBL) group tutors at two medical schools.

1.4 Correlations were run between SJT scores (using raw SJT scores and also scaled SJT scores) and in-training performance variables created from the questionnaire response data.

1.5 Statistically significant positive correlations were found between in-training performance and SJT scores for the School A PBL group at the total level (most importantly) and also at the Perspective-taking and Team Involvement domain levels. These are encouraging findings.

1.6 An association between SJT scores and in-training performance was not found at School B or for the GP tutor groups at School A. Possible reasons for this are discussed.

1.7 This study provides some positive early evidence to support the validity of the UKCAT SJT.

1.8 There is evidence to suggest that an Item Response Theoretic (IRT) approach that requires a uni-dimensional factor structure is not a suitable calibration method for SJTs.

1.9 Further work is required to finalise suitable outcome measure(s) for the validation of non-cognitive attributes in this context.

1.10 Strengths, limitations and points of learning regarding the present study are summarised and important considerations for future validation work are discussed.

2. Context

2.1 A situational judgement test (SJT) was piloted in 2012 and introduced live in 2013 to evaluate important non-academic attributes as part of the UKCAT. The use of SJTs in medical and dental selection is widening and, following role analyses of numerous medical specialities, it is widely acknowledged that non-academic attributes are essential requirements for a doctor or dentist.

2.2 As with any assessment, it is essential to validate the SJT with regards to linking SJT results to in-training performance. As well as ensuring the defensibility of the new approach, the outcomes can also be used to identify areas for further improvement and development as well as informing policy to optimise effectiveness and efficiency of the selection system in future.

2.3 A robust predictive validation study requires a longitudinal methodology. However, the first cohort of students, who sat the SJT pilot in 2012, began their first year of medical/dental training in autumn 2013. As such, the present study took the opportunity to conduct a preliminary validation study, focusing on this cohort of students.

2.4 The key objectives were to

- better understand the ways in which a validation study can be suitably conducted in this context
- identify points of learning to inform a specification for a longitudinal validation study
- understand its applicability to the broad range of schools within the UKCAT Consortium
- provide an initial indication of the relationship between SJT scores and in-training performance at this early stage in the SJT implementation process

3. Methodology

3.1. The cohort of first year medical/students who began their courses in autumn 2013, and sat the SJT pilot in 2012, were invited to participate in this study. Two medical schools participated.

3.2. Both raw and scaled pilot SJT scores were used as separate predictor variable. The raw scores comprise of a fuller range of items and therefore it is very likely that they benefit from a greater degree of content validity. Scaled scored, however, mitigate the non-equivalence (minor variation in difficulty) of the un-equated raw scores. The development of the scaled scores, and their potential limitation in this context, is described below. Further detail regarding the scaling of pilot data is available within the IRT Technical Appendix report (WPG, February 2013).

3.3. In-training performance ratings, provided by tutors in response to a bespoke questionnaire, were used as the outcome variable. An established methodology was used (see Patterson et al, 2005)[1], using the behavioural performance indicators from the UKCAT Role Analysis for Medical and Dental Students. Questionnaires were issued to GP and GP Community tutors at School A and to Problem Based Learning (PBL) group tutors at both schools.

3.4. Both medical schools received ethical approval to participate in the research. All potential participants were briefed about the research. It was explained that at no point would participants' names be known to WPG or reported in any way and that they were free to not participate or withdraw. All data were held securely with password encryption where relevant.

4. Results

4.1. 432 in-training performance questionnaires were able to be matched to SJT scores. 101 were from School A GP tutors, 138 from School A PBL tutors, 79 from School A GP Community tutors and 114 were from School B PBL tutors.

4.2. Data related to 310 students (51% female, 47% White, 84% British, mean age 19 years).

[1] Patterson F, Carr V, Plint S. (2005). *General practice competency validation exercise: technical report to GP National Recruitment Office*. Sollihull: National Recruitment Office for GP Training, 2005.

4.3. Each of the 21 questionnaire items, plus the three 'overall' ratings (one per domain) created 24 immediate variables for analysis. Further to this, *mean score variables* were created for:

- each of the three domains (referred to as *Integrity mean, PT mean* and *TI mean*)
- all 21 single items, not including the three 'overall' ratings (referred to as *Total mean* – this is the single most meaningful outcome variable)
- the three overall ratings (referred to as *Overall mean*)

4.4. Data cleaning was conducted according to robust statistical protocols.

4.5. Table 1 below shows the descriptive statistics for outcome variables and SJT scores for the PBL groups. Table 2 shows this for the GP and GP Community groups (School A only). The maximum possible SJT raw score ranged from 252 to 272 depending on which form was sat. The minimum possible raw score was 0. SJT scale scores were scaled to a mean of 0 and a standard deviation of 0.45.

4.6. T tests revealed that the differences in the key outcome variables between the two schools were either significant or approaching significance, with School A participants receiving higher performance ratings than those at School B.

Table 1: Descriptive statistics to compare outcome variables for School A and School B PBL groups

		N	Mean	Std. Deviation	Min score	Max score
Integrity mean*	School A	108	5.09	0.80	2.75	6.00
	School B	94	4.82	0.63	3.25	6.00
PT mean	School A	124	4.99	0.84	2.33	6.00
	School B	94	4.77	0.77	2.71	6.00
TI mean*	School A	125	4.87	0.96	1.00	6.00
	School B	94	4.54	1.00	1.17	6.00
Total mean	School A	125	4.93	0.87	2.27	6.00
	School B	94	4.71	0.76	2.75	6.00
Overall mean	School A	123	4.97	0.86	2.00	6.00
	School B	94	4.74	0.83	2.67	6.00
SJT raw	School A	124	205.51	12.94	166.00	236.00
	School B	94	207.28	12.53	165.00	240.00
SJT scaled*	School A	125	0.04	0.42	-1.12	1.18
	School B	94	0.15	0.36	-0.87	1.01

* difference is significant at the $p < .05$ level

Table 2: Descriptive statistics for School A GP and GP Community

	N	Mean	SD	Min	Max		N	Mean	SD	Min	Max
	GP						GP Community				
Integrity mean	95	5.03	.52	3.25	6.00	Integrity mean	75	5.05	.64	3.38	6.00
PT mean	94	5.00	.68	2.57	6.00	PT mean	75	5.08	.68	3.29	6.00
TI mean	95	4.94	.75	1.67	6.00	TI mean	75	4.88	.77	2.67	6.00

Total mean	95	4.99	.61	2.50	6.00	Total mean	75	5.00	.64	3.36	6.00
Overall mean	95	5.04	.64	2.50	6.00	Overall mean	75	5.05	.66	3.67	6.00

4.7. Internal consistency reliability of the in-training performance questionnaire was in the highly acceptable range for all sub scales for all groups at both schools (Cronbach alpha .83 to .99).

4.8. Pearson's correlations (two-tailed) were used to examine the degree of association between SJT raw scores and the key outcome variables for each of the four samples. These are shown in Tables 3a and 3b below.

4.9. **The SJT raw score correlates significantly and positively with performance scores for the School A PBL group**. This is with the Total mean outcome variable (the single most meaningful outcome variable) ($r = .27$, $p < .01$), at the domain level for Perspective Taking ($r = .28$, $p < .01$) and Team Involvement ($r = .26$, $p < .01$) and finally with the 'overall' mean ($r = .19$, $p < .05$).

Table 3a: Correlations between predictor and outcome measures per school/tutor group

		Integrity mean	PT mean	TI mean	Total mean	Overall mean
SJT raw score						
School A GP	Pearson *r*	-0.16	-0.17	-0.14	-0.17	-0.18
	N	95	94	95	95	95
School A PBL	Pearson *r*	0.18	**0.28****	**0.26****	**0.27****	**0.19***
	N	107	123	124	124	122
School A Community	Pearson *r*	-0.10	-0.02	-0.02	-0.05	0.08
	N	75	75	75	75	75
School B PBL	Pearson *r*	-0.01	-0.04	-0.00	0.01	-0.01
	N	94	94	94	94	94

* correlation is significant at the $p < .05$ level
** correlation is significant at the $p < .01$ level

Table 3b: Correlations between scaled scores and outcome measures per school/tutor group

		Integrity mean	PT mean	TI mean	Total mean	Overall mean
SJT scaled score						
School A GP	Pearson *r*	-0.04	-0.04	-0.09	-0.06	-0.05
	N	95	94	95	95	95
School A PBL	Pearson *r*	0.10	0.14	**0.20***	**0.20***	0.14
	N	107	123	124	124	122
School A Community	Pearson *r*	0.01	0.07	0.04	0.04	-0.02
	N	75	75	75	75	75
School B PBL	Pearson *r*	0.04	0.10	0.03	0.06	0.01
	N	94	94	94	94	94

* correlation is significant at the $p < .05$ level
** correlation is significant at the $p < .01$ level

4.10. Additionally, Table 3b shows that the SJT scaled score correlates significantly and positively with total in-training performance scores for the School A PBL group ($r = .20$, $p < .05$) and also with the Team Involvement mean score ($r = .20$, $p < .05$).

4.11. These are encouraging findings and indicate early SJT validation evidence in terms of criterion-matched in-training performance as rating by School A PBL group tutors.

4.12. However this relationship was found at School B. This is discussed further below.

5. Conclusions

5.1. This study highlights some important implications for a full longitudinal SJT validation study, and some encouraging early validation evidence at one of the two participating medical schools. Interpretation of results raises a number of considerations.

5.2. Research leads at School A and School B reported differences in the ways in which data were gathered, owing to the different PBL teaching structures. Firstly, at School A, PBL tutors were instructed face to face. They were talked through the purposes of the study, the guidance documentation, how to use the rating scale, and it was emphasised that tutors should seek to use the full scale where appropriate. At School B this information was issued to tutors remotely and the information about the value of the project and the way in which to use the rating scale was not built upon beyond the written guidance. Hence the approach taken at School A may have encouraged greater engagement from tutors and perhaps a typically more discerning approach. This may explain the greater SD's for School A compared to School B as shown in Table 1.

5.3. Another explanation for this could also be that the School B tutors were also responsible for the pastoral care of the students, whereas this was not the nature of the tutor/student relationship at School A. As was suggested, this may have inhibited the School B tutors from awarding very low ratings, whereas at School A this does not appear to have been a problem.

5.4. Additional considerations include the difference in sample size between the two PBL samples, and as previously explained the difference in the number of tutor/student interactions. School A data is of better quality in this respect because the sample size is larger and tutors' judgements are based on a greater, and hence more reliable, number of interactions.

5.5. Secondly, it is interesting that the GP and GP Community ratings show either negligible or negative correlations (not significant however) with the SJT. This was not expected. This could be partly due to the fact that guidance was distributed and the ratings were undertaken remotely. It is possible that, as with School B as previously discussed, this approach led to a less discerning approach from GP and GP Community tutors compared to the PBL tutors.

5.6. It is interesting that many of the School A PBL tutors annotated 'not seen' or 'not observed' when they didn't answer an Integrity question. It seems that observation of Integrity behaviours within a PBL environment may be particularly problematic.

5.7. Project leads at the medical school suggested that students who declined to take part in the study tended to be the lower performing students. Perhaps this means that the requirement of consent limited the distribution of scores at both medical schools.

5.8. It is very important to conduct future validation research that can benefit from more robust predictor and outcome data. The present study offers a helpful model to build upon.

5.9. There are indications that the in-training performance questionnaire is an appropriate approach per se. It would be helpful to discuss the validity of this in greater detail with stakeholders. This would help to develop our understanding of how well each of the items translates to the various teaching contexts, and the points where these become sufficiently valid and reliable indicators during the students' training programme.

5.10. Peer ratings may be a suitable additional source of data if opportunity to demonstrate the outcome criterion to tutors is particularly problematic in some schools/contexts.

5.11. The differences observed between the two medical schools suggest that a one-size national validation is not an appropriate approach. Stakeholders considered from the outset that this may be the case. However further discussions would be useful in order to unpick the differences and similarities in terms of the delivery of teaching across locations and types of schools. This will help to establish clearer research hypotheses.

5.12. Future validation will be able to use SJT data that is from operational use rather than from piloting.

5.13. It is also important to note that the positive correlations found with in-training data were notably larger with the raw score as opposed to the scaled score. This implies that despite the non-equivalence challenges with the raw scores as explained previously, they were more predictive of in-training performance than the scaled scores. An explanation for this finding could be due to the restricted validity of the scaled scores. To reiterate, items had to be dropped from the scale score calibration if they did not conform sufficiently to the uni-dimensionality requirements of the IRT approach. This has important implications for the pre-calibration of the UKCAT SJT and also for SJT development more generally. Present evidence suggests that it is problematic to restrict the broad and richly contextual content validity of SJTs to a factor structure that is uni-dimensional.

5.14. It will be vital to engage medical and dental schools within the UKCAT Consortium and secure the participation of approximately five schools in a longitudinal predictive validation study.

5.15. In summary, the present study offers some positive early findings and has unearthed some considerations that warrant further exploration to inform the design of a longer term and wider reaching validation study.