Koczwara, A., Patterson, F., Zibarras, L. D., Kerrin, M., Irish, B. & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. Medical Education, 46(4), pp. 399-408. doi: 10.1111/j.1365-2923.2011.04195.x

CITY UNIVERSITY LONDON

EST 1894

City Research Online

Original citation: Koczwara, A., Patterson, F., Zibarras, L. D., Kerrin, M., Irish, B. & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. Medical Education, 46(4), pp. 399-408. doi: 10.1111/j.1365-2923.2011.04195.x

Permanent City Research Online URL: http://openaccess.city.ac.uk/12509/

Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection

Anna Koczwara,[1–3] Fiona Patterson,[1–3] Lara Zibarras,[1–3] Maire Kerrin,[1–3] Bill Irish[1–3] & Martin Wilkinson[1–3]

OBJECTIVES This study aimed to evaluate the validity and utility of and candidate reactions towards cognitive ability tests, and current selection methods, including a clinical problem-solving test (CPST) and a situational judgement test (SJT), for postgraduate selection.

METHODS This was an exploratory, longitudinal study to evaluate the validities of two cognitive ability tests (measuring general intelligence) compared with current selection tests, including a CPST and an SJT, in predicting performance at a subsequent selection centre (SC). Candidate reactions were evaluated immediately after test administration to examine face validity. Data were collected from candidates applying for entry into training in UK general practice (GP) during the 2009 recruitment process. Participants were junior doctors (n = 260). The mean age of participants was 30.9 years and 53.1% were female. Outcome measures were participants' scores on three job simulation exercises at the SC.

RESULTS Findings indicate that all tests measure overlapping constructs. Both the CPST and SJT independently predicted more variance than the cognitive ability test measuring non-verbal mental ability. The other cognitive ability test (measuring verbal, numerical and diagrammatic reasoning) had a predictive value similar to that of the CPST and added significant incremental validity in predicting performance on job simulations in an SC. The best single predictor of performance at the SC was the SJT. Candidate reactions were more positive towards the CPST and SJT than the cognitive ability tests.

CONCLUSIONS In terms of operational validity and candidate acceptance, the combination of the current CPST and SJT proved to be the most effective administration of tests in predicting selection outcomes. In terms of construct validity, the SJT measures procedural knowledge in addition to aspects of declarative knowledge and fluid abilities and is the best single predictor of performance in the SC. Further research should consider the validity of the tests in this study in predicting subsequent performance in training.

[1]Work Psychology Group, Department of Xxx, University of Cambridge, Cambridge, UK
[2]Department of Psychology, City University London, London, UK
[3]Work Psychology Group, Severn Deanery, West Midlands Deanery, Xxx, UK

Correspondence: Dr Lara Zibarras, Department of Psychology, City University London, Northampton Square, London EC1V 0HB, UK. Tel: 00 44 20 7040 4573; Fax: 00 44 20 7040 8581; E-mail: lara.zibarras.1@city.ac.uk, l.zibarras@city.ac.uk

Dispatch: 5.12.11	Journal: MEDU	CE: K. Karthik
Author Received:	No. of pages: 10	PE: Prasanna

B

MEDU — **Journal Name** — **4 1 9 5** — **Manuscript No.**

INTRODUCTION

Large-scale meta-analytic studies show that general cognitive ability tests are good predictors of job performance across a broad range of professions and occupational settings.[1–4] Cognitive ability tests assess general intelligence (IQ) and have been used for selection in high-stakes contexts such as military and pilot selection.[5,6] Cognitive ability tests have been used and validated for medical school admissions,[7] but not yet for selection into postgraduate medical training. Previous research focused on medical school admissions; the use of cognitive ability tests in medical selection remains controversial.[7–9] This paper presents an evaluation of the validity and utility of cognitive ability tests as a selection methodology for entry into training in UK general practice (GP). The selection methodology currently used for entry into GP training demonstrates good evidence of reliability, and face and criterion-related validity.[10–13] The selection methodology comprises three stages: eligibility checks are succeeded by shortlisting via the completion of two invigilated, machine-marked tests, and this is followed by the administration of a clinical problem-solving test (CPST) and a situational judgement test (SJT).[10,11] The CPST requires candidates to apply clinical knowledge to solve problems involving a diagnostic process or a management strategy for a patient. The SJT focuses on a variety of non-cognitive professional attributes (empathy, integrity, resilience) and presents candidates with work-related scenarios in which they are required to choose an appropriate response from a list. The final stage of selection comprises a previously validated selection centre (SC) using three high-fidelity job simulations which assess candidates over a 90-minute period, involving three separate assessors and an actor. These are: (i) a group discussion exercise referring to a work-related issue; (ii) a simulated patient consultation in which the candidate plays the role of doctor and an actor plays the patient, and (iii) a written exercise in which the candidate prioritises a list of work-related issues and justifies his or her choices.[10,14]

In this study, two cognitive ability tests were piloted alongside the live selection process for 2009 in order to explore their potential for use in future postgraduate selection. Cognitive ability tests were considered here because if either cognitive ability test were to demonstrate improved validity over and above that of the existing CPST and SJT short-listing tests, this might indicate potential for significant gains in the effectiveness and efficiency of the current process. For example, cognitive ability tests might demonstrate improved utility because they are shorter (and thus take less candidate time) and are not specifically designed for GP selection and would therefore not require clinicians' time during development phases. The first cognitive ability test was the Ravens Advanced Progressive Matrices.[15] This is a power test focusing on non-verbal mental ability (referred to as NVMA in the present study) and measures general fluid intelligence, including observation skills, clear thinking ability, intellectual capacity and intellectual efficiency. The test takes 40 minutes to complete. The NVMA score indicates the candidate's potential for success in high-level positions that require clear, accurate thinking, problem identification and evaluation of solutions.[15] These are abilities shown to be relevant to GP training.[16] The second cognitive ability test was a speed test designated the Swift Analysis Aptitude Test.[17] It is a cognitive ability test battery (referred to as CATB in the present study) and consists of three short sub-tests measuring verbal, numerical and diagrammatic analysis abilities. This test allows three specific cognitive abilities to be measured in a relatively short amount of time, giving an overall indication of general cognitive ability. The whole CATB takes 18 minutes and therefore might offer practical utility compared with other IQ test batteries in which single test subsets typically take 20–30 minutes.[17] Both the NVMA and CATB have good internal reliability and have been validated for selection purposes in general professional occupations,[15,17] but not in selection for postgraduate training in medicine.

This present study examines the validity of two forms of cognitive ability test and the CPST and SJT selection tests in predicting performance in the subsequent SC. Table 1 outlines each of the tests evaluated in the present study. Example items are given in Table S1 (online). Theoretically, the NVMA **10** is a measure of fluid intelligence; it measures the ability to logically reason and solve problems in new situations, independent of learned knowledge.[18] Similarly, the CATB measures fluid intelligence, but also measures elements of crystallised intelligence (experience-based ability in which knowledge and skills are acquired through educational and cultural experience)[17] because a level of procedural knowledge regarding verbal and numerical reasoning is required to understand individual items. The CPST measures crystallised intelligence, especially declarative knowledge, and is designed as a test of attainment examining learned knowledge gained through previous medical training. Finally, the SJT is a measure designed to test non-cognitive professional attributes.

Table 1 Description of the tests and outcome measures used in the study (Table S1 [online] gives examples of items in the tests) **29**

Test name	Theoretical underpinning of test	Description
Ravens Advanced Progressive Matrices	Power test measuring fluid intelligence Power tests generally do not impose a time limit on completion The non-verbal format reduces culture bias	The advanced matrices differentiate between candidates at the high end of ability The test has 23 items, each with 8 response options; thus a maximum of 23 points are available (1 point per correct answer)
Swift Analysis Aptitude	Speed test measuring fluid intelligence and some aspects of crystallised intelligence Speed tests focus on the amount of questions answered correctly within a specific timeframe	Three subsets, each with 8 items with 3–5 response options; thus a maximum of 24 points are available (1 point per correct answer)
Clinical problem-solving test (CPST)	Clinical problem-solving test measuring crystallised abilities, especially declarative knowledge It is designed as a power test	The CPST has 100 items to be completed within 90 minutes
Situational judgement test (SJT)	Test designed to measure non-cognitive professional attributes beyond clinical knowledge It is designed as a power test	The SJT has 50 items to be completed in 90 minutes There are two different types of items: ranking and choice
Selection centre	Multitrait–multimethod assessment centre in which candidates are rated on their observed behaviours in three exercises	(i) A group exercise, involving a group discussion referring to a work-related issue (ii) A simulated patient consultation in which the candidate plays the role of the doctor and an actor plays the patient (iii) A written exercise in which candidates prioritise a list of work-related issues and justify their choices

In making judgements about response options, candidates need to apply relevant knowledge, skills and abilities to resolve the issues they are presented with.[19,20]

We would anticipate some relationship among all four tests as they measure overlapping constructs. Because the two cognitive ability tests both measure similar constructs, we might expect a reasonably strong correlation between the two. The CPST, although essentially a measure of crystallised intelligence, is likely to also entail an element of fluid intelligence as verbal reasoning is necessary to understand the situations presented in the question items. Thus, we would expect the CPST to be positively related to the cognitive ability tests. The construct validity of SJTs is less well known; research suggests that they may relate to both cognitive ability[21] and learned job knowledge.[22] Indeed, a meta-analysis showed that SJTs had an average correlation of $r = 0.46$ with cognitive ability.[20] In the present context, although the SJT was designed to measure non-cognitive domains, we may expect a positive association between the SJT and the two cognitive ability tests. This is because theory suggests that intelligent people may learn more quickly about the non-cognitive traits that are more effective in the work-related situations described in the SJT.[23]

The present study aimed to evaluate the construct, predictive, incremental and face validity of two cognitive ability tests in comparison with the present CPST and SJT selection tests. In examining predictive and incremental validity, we used a previously validated approach[11] with overall performance at the SC as an outcome measure. Performance at the SC has been linked to subsequent training performance[14] and predicts supervisor ratings 12 months into training.[24] We therefore posed the following four research questions:

1   Construct validity: what are the inter-correlations among the two cognitive ability tests, the CPST and the SJT?

2 Predictive validity: do scores on each of the tests independently predict subsequent performance at the SC?
3 Incremental validity: compared with the current CPST and SJT, do the cognitive ability tests (NVMA and CATB) each account for additional variance in performance at the SC?
4 Face validity: do candidates perceive the tests to be fair and appropriate?

METHODS

Design and procedure

Data were collected during the 2009 recruitment process for GP training in the West Midlands region in the UK. Candidates were invited to participate on a voluntary basis and gave consent for their scores at the SC to be accessed. It was emphasised that all data would be used for evaluation purposes only. Candidates successful at shortlisting were invited to the SC, where their performance on job simulations forms the basis for job offers. For each of the three SC exercises, assessors rated candidates' performance around four of the following competencies: Empathy and Sensitivity; Communication Skills; Coping with Pressure, and Problem Solving and Professional Integrity. These competencies were derived from the previous job analysis[16] and assessors were provided with a 4-point rating scale (1 = poor, 4 = excellent) with which to rate the candidate and behavioural anchors to assist in determining the rating. For each of the exercises, the sum of the ratings for the four competencies was calculated to create a total score. Exercise scores were summed to create the overall SC score.

Associations among all variables in the study were examined using Pearson correlation coefficients; note that none of the correlations reported in the results have been corrected for attenuation. To investigate the relative predictive abilities of the four tests in the study (two cognitive ability tests and the current selection methods, CPST and SJT), two types of regression analyses were conducted using overall performance at the SC as the dependent variable. Firstly, we used hierarchical regression analysis. With this method, the predictor variables are added into the regression equation in an order determined by the researcher; thus in the present context, the CPST and SJT predictors would be added in the first step, and then the cognitive ability test(s) would be added in the second step to determine the additional variance in the SC score predicted by the cognitive ability test. Secondly, we used stepwise regression analysis. With this method, the order of relevance of predictor variables is identified by the statistical package (SPSS Version X.XX; SPSS, Inc., Chicago, IL, 11 USA) to establish which predictors independently predict the most variance in SC scores.

Sampling

A total of 260 candidates agreed to participate in the study. Of these, 53.1% were female. Their mean age was 30.9 years (range: 24–54 years). Participants reported the following ethnic origins: White (38%); Asian (43%); Black (12%); Chinese (2%); Mixed (2%), and Other (3%). The final sample of candidates for whom NVMA data were available numbered 219 because 26 candidates did not consent to the matching of their pilot data with live selection data and a further 15 candidates were not invited to the SC. The final sample of candidates for whom CATB data were available numbered 188 because, of the 215 candidates who initially completed the CATB, 13 consented to participate in the pilot but did not want their pilot data matched to their live selection data, and a further 14 candidates did not pass the shortlisting stage and so were not invited to the SC. There were no significant demographic differences between the two samples and the overall demographics of both were similar to those of the live 2009 candidate cohort.

RESULTS

Table 2 provides the descriptive statistics for all tests for the full pilot sample ($n = 260$). The means and ranges for both cognitive ability tests were similar to those typically found in their respective comparison norm groups, which comprised managers, professional and graduate-level employees; thus the sample's cognitive ability test scores were comparable with those of the general population.[15,17] The CPST and SJT scores, the two cognitive ability tests and the SC data (including all exercise scores and overall scores) were normally distributed.

What are the inter-correlations among the two cognitive ability tests, the CPST and the SJT?

Significant positive correlations were found between the NVMA and the CATB ($r = 0.46$), CPST ($r = 0.36$) and SJT ($r = 0.34$), and also between the CATB and CPST ($r = 0.41$) and SJT ($r = 0.47$) (all $p < 0.01$) (Table 3).Thus, the cognitive ability tests have both

Table 2 Descriptive statistics and correlations among study variables

	n	Mean	SD	Range	1	2	3	4	5	6	7	8
NVMA	234	11.12	4.07	1–21	(0.85)							
CATB	202	12.38	4.23	2–24	0.46†	(0.76)						
CPST	202	254.62	39.49	99–326	0.36†	0.41†	(0.86)					
SJT	202	254.23	39.20	110–331	0.34†	0.47†	0.44†	(0.85)				
Group exercise	188	12.68	2.67	4–16	0.19†	0.28†	0.28†	0.39†	(0.90)			
Written exercise	188	12.69	2.43	5–16	0.15*	0.23†	0.19†	0.28†	0.30†	(0.89)		
Simulation exercise	188	12.84	2.92	5–16	0.29†	0.32†	0.32†	0.40†	0.28†	0.20†	(0.92)	
SC overall score	188	38.21	5.72	16–48	0.30†	0.39†	0.38†	0.50†	0.74†	0.67†	0.73†	(0.87)

$* \ p < 0.05; \ † \ p < 0.01$ (two-tailed)

Correlations between variables were uncorrected for restriction for range. Numbers in parentheses are the reliabilities for the selection methods for the overall 2009 recruitment round. For the NVMA and CATB, these reliabilities are those reported in the respective manuals

SD = standard deviation; NVMA = non-verbal cognitive ability test; CATB = cognitive ability test battery; CPST = clinical problem-solving test; SJT = situational judgement test; SC = selection centre

Table 3 Regression analyses

	NVMA dataset, n = 219				CATB dataset, n = 188		
	B	SE B	β		B	SE B	β
Hierarchical regression analysis							
Step 1, $R^2 = 0.31$				Step 1, $R^2 = 0.29$			
Constant	13.77	2.51		Constant	13.63	2.97	
SJT	0.07	0.01	0.49*	SJT	0.07	0.01	0.42*
CPST	0.03	0.01	0.18†	CPST	0.03	0.01	0.20†
Step 2, $\Delta R^2 = 0.01$				Step 2, $\Delta R^2 = 0.02$			
NVMA	0.15	0.09	0.10	CATB	0.21	0.10	0.15‡
Stepwise regression analysis							
Step 1, $R^2 = 0.29$				Step 1, $R^2 = 0.26$			
Constant	17.33	2.19		Constant	18.26	2.53	
SJT	0.08	0.01	0.54*	SJT	0.08	0.01	0.51*
Step 2, $\Delta R^2 = 0.01$				Step 2, $\Delta R^2 = 0.03$			
CPST	0.03	0.01	0.18†	CPST	0.03	0.01	0.20†
				Step 2, $\Delta R^2 = 0.02$			
				CATB	0.21	0.10	0.15‡

$* \ p < 0.001, \ † \ p < 0.01, \ ‡ \ p < 0.05$

SE = standard error; NVMA = non-verbal cognitive ability test; CATB = cognitive ability test battery; CPST = clinical problem-solving test; SJT = situational judgement test

common and independent variance with the CPST and SJT and to some extent measure overlapping constructs.

Do scores on each of the tests independently predict subsequent performance at the SC?

The analyses in Table 2 showed a positive correlation between NVMA scores and all SC exercises (group, $r = 0.19$; simulation, $r = 0.29$ [both $p < 0.01$]; written, $r = 0.15$ [$p < 0.05$]). However, in comparison with the CPST and SJT, both the CPST and SJT had substantially higher correlations with the three SC exercises (CPST, $r = 0.19$–0.32; SJT, $r = 0.28$–0.40 [all $p < 0.01$]). Furthermore, both the CPST and SJT had higher correlations with overall SC scores ($r = 0.38$ and $r = 0.50$, respectively) compared with the NVMA ($r = 0.30$) (all $p < 0.01$).

Results show a positive correlation between the CATB and the SC exercises (group, $r = 0.28$; simulation, $r = 0.32$; written, $r = 0.23$ [all $p < 0.01$]). The CPST correlated with the group and simulation exercises to the same extent as the CATB ($r = 0.28$ and $r = 0.32$, respectively [both $p < 0.01$]), but had a lower correlation with the written exercise ($r = 0.19$ [$p < 0.01$]) (Table 2). The SJT had higher correlations with all exercises (group, $r = 0.39$; written, $r = 0.28$; simulation, $r = 0.40$ [all $p < 0.01$]). Further, the CPST had a lower correlation with overall SC scores compared with the CATB ($r = 0.38$ and $r = 0.39$, respectively [both $p < 0.01$]), but the SJT had a higher correlation ($r = 0.50$ [$p < 0.01$]).

Thus, overall findings indicate that, of all the tests, the SJT had the highest correlations with performance at the SC. The SJT and CPST were more effective predictors of subsequent performance than the NVMA; the SJT was a better predictor of performance at the SC than the CATB, and the CPST had a similar predictive value to the CATB.

Compared with the current CPST and SJT, do the cognitive ability tests (NVMA and CATB) each account for additional variance in performance at the SC?

We established the extent to which the cognitive ability tests each accounted for additional variance above the current CPST and SJT. To examine the extent to which NVMA scores predicted overall SC scores over and above the CPST and SJT scores combined, a hierarchical multiple regression was performed (Table 3). Scores on the CPST and SJT were entered in the first step, which explained 31.3%

of the variance in SC overall score ($R^2 = 0.31$, $F_{(2,216)} = 49.24$; $p < 0.001$); however, adding the NVMA in the second step offered no unique variance over the CPST and SJT ($\Delta R^2 = 0.01$, $F_{(2,215)} = 2.85$; $p = 0.09$). A stepwise regression was also performed (Table 3) to establish which tests independently predicted the most variance in SC scores. Scores on the SJT were entered into the first step, indicating that the SJT explains the most variance (28.9%) of all the tests ($R^2 = 0.29$, $F_{(1,217)} = 88.02$; $p < 0.001$). Scores on the CPST were entered into the second and final step, explaining an additional 2.5% of the variance ($\Delta R^2 = 0.03$, $F_{(2,216)} = 7.73$; $p = 0.01$). The NVMA was not entered into the model at all, confirming its lack of incremental validity over the two current short-listing assessments.

To establish the extent to which the CATB predicted SC performance over and above both the CPST and SJT, a hierarchical multiple regression was performed (Table 3), repeating the method described above. Scores on the SJT and CPST were entered into the first step and explained 28.6% of the variance ($R^2 = 0.29$, $F_{(2,185)} = 36.99$; $p < 0.001$); entering the CATB into the second step explained an additional 1.7% of the variance in overall SC performance ($\Delta R^2 = 0.02$, $F_{(1,184)} = 4.44$; $p = 0.04$). A stepwise regression was also performed, entering SJT scores into the first step. This explained the most variance in SC performance (25.5%) of all the tests ($R^2 = 0.26$, $F_{(1,186)} = 63.53$; $p < 0.001$). Scores on the CPST were entered into the second step, explaining an additional 3.1% of the variance ($\Delta R^2 = 0.03$, $F_{(1,185)} = 8.05$; $p = 0.005$). Finally, scores on the CATB were entered into the final step, explaining an additional 1.7% of the variance ($\Delta R^2 = 0.02$, $F_{(1,184)} = 4.44$; $p = 0.04$). Overall, findings indicate that the CATB does add some incremental validity in predicting SC performance.

A stepwise regression was also performed to establish which of the four tests (SJT, CPST, NVMA, CATB) independently predicted the most variance in SC scores. Scores on the SJT were entered into the first step, indicating that the SJT explains the most variance (25.5%) of all the tests ($R^2 = 0.26$, $F_{(1,186)} = 63.53$; $p < 0.001$). Scores on the CPST were entered into the second step, explaining an additional 3.1% of the variance ($\Delta R^2 = 0.03$, $F_{(1,185)} = 8.05$; $p = 0.005$). Finally, CATB scores were entered into the third step, explaining an additional 1.7% of the variance ($\Delta R^2 = 0.02$, $F_{(1,184)} = 4.44$; $p = 0.04$). The NVMA was not entered into the model at all and thus lacked incremental validity over the other tests. Overall, findings indicate that the NVMA

does not add incremental validity in predicting SC performance, but the CATB does add some incremental validity.

Do candidates perceive the tests as fair and appropriate?

All participants were asked to complete a previously validated candidate evaluation questionnaire,[12,25] based on procedural justice theory, regarding their perceptions of the tests. A total of 249 candidates completed the questionnaire (96% response rate), in which they indicated their level of agreement with several statements regarding the content of each test. These results are shown in Table 4, along with feedback on the SJT and CPST from the live 2009 selection process. Overall results show that the CPST received the most favourable feedback, followed by the SJT. Candidates did not react favourably to the NVMA, mainly as a result of perceptions of low job relevance and insufficient opportunity to demonstrate ability. The CATB was also relatively negatively received; feedback was slightly better than for the NVMA but still markedly worse than for the CPST and SJT. This notably reflected perceptions of the CATB as providing insufficient opportunity to demonstrate

ability and as not helping selectors to differentiate among candidates.

DISCUSSION

Two cognitive ability tests were evaluated as potential selection methods for use in postgraduate selection and were compared with the current selection tests, the CPST and SJT. Results show positive and significant relationships among the CPST and SJT and cognitive ability tests, indicating that they measure overlapping constructs. For both the CPST and the SJT, the correlation with the CATB was higher than with the NVMA. This is probably because the CATB, CPST and SJT are all presented in a verbal format, whereas the NVMA has no verbal element and is presented in a non-verbal format.

Implications for operational validity

Considering both the predictive validation analyses and candidate reactions, the CPST and SJT (measuring clinical knowledge and non-cognitive professional attributes, respectively) have been shown to be a better combination in predicting SC performance

*Table 4 Procedural justice reactions to the tests*

	NVMA (% of candidates, n = 249)			CATB (% of candidates, n = 195)			CPST (% of candidates, n = 2947)			SJT (% of candidates, n = 2947)		
	SD/D	N	A/SA	SD/D	N	A/SA	SD/D	N	A/SA	SD/D	N	A/SA
The content of the test was clearly relevant to GP training	**62**	23	16	**47**	29.2	24	3	7	**89**	13	22	**63**
The content of the test seemed appropriate for the entry level I am applying for	**40**	33	26	32	**35**	33	4	9	**85**	9	22	**68**
The content of the test appeared to be fair	33	31	**37**	30	29	**41**	4	10	**85**	19	27	**53**
The test gave me sufficient opportunity to indicate my ability for GP training	**66**	23	11	**58**	26	15	9	18	**72**	36	28.9	**34**
The test would help selectors to differentiate between candidates	**57**	21	20	**50**	25	24	10	20	**67**	35	29.1	**34**

NVMA = non-verbal cognitive ability test; CATB = cognitive ability test battery; CPST = clinical problem-solving test; SJT = situational judgement test; SD/D = strongly disagree or disagree; N = neither agree nor disagree; A/SA = agree or strongly agree; GP = general practice

compared with the cognitive ability tests. The NVMA results showed a moderate correlation with SC performance, but no incremental validity over the CPST and SJT. Furthermore, the NVMA was negatively received by candidates. The CATB was moderately correlated with SC performance and showed incremental validity over and above the CPST and SJT; however, the CPST and SJT in combination demonstrated significantly more variance in SC performance than when either test was combined with the CATB. The results show that the test measuring non-cognitive professional attributes (the SJT) is the best single predictor of subsequent performance at the SC. For operational validity, the best combination of tests in terms of explaining the greatest amount of variance in SC performance included the CPST, SJT and CATB. However, the increase in variance explained by the CATB is not large and has to be weighed against the cost and time implications of increasing the amount of test-taking time per candidate.

Theoretical implications

As the SJT was the best single predictor of SC performance, it could be argued that the constructs that best predict subsequent performance in job simulations include a combination of crystallised and fluid intelligence, along with 'non-cognitive' professional attributes measured by the SJT. The construct validity of SJTs has been a subject for debate amongst researchers and we argue that the results presented here provide further insights. Motowidlo and Beier[23] suggest that the procedural knowledge measured by an SJT includes implicit trait policies, which are implicit beliefs regarding the costs and benefits of how personality is expressed and its effectiveness in specific jobs (which is likely to relate to the way in which professional attributes are expressed in a work context). Results suggest that the SJT broadens the constructs being measured (beyond declarative knowledge and fluid intelligence) and therefore the SJT demonstrates the highest criterion-related validity in predicting performance in high-fidelity job simulations in which a range of different work-related behaviours are measured (beyond knowledge and cognitive abilities). However, in order to build and extend the current research and to test the ideas presented by Motowidlo and Beier,[23] we recommend that future research exploring the construct validity of the SJT should also include a measure of personality and implicit trait policies to test this possible association. We therefore present Fig. 1, a diagram illustrating potential pathways among variables that could be considered in future research.
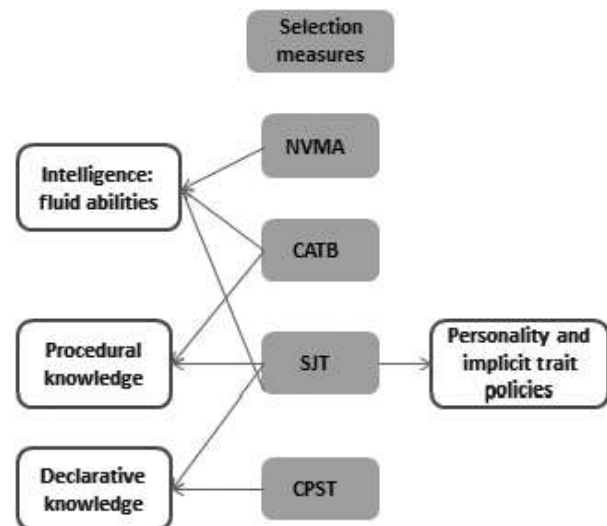
Figure 1 Selection measures and their hypothesised construct validity. NVMA = non-verbal mental ability test; CATB = cognitive ability test battery; SJT = situational judgement test; CPST = clinical problem-solving test

Practical implications

Although the CATB may appear to offer some advantages relating to cost savings in terms of administration and test development, there are several reasons why replacing the CPST with the CATB might have negative implications in this context. The first reason relates to patient safety: using the general cognitive ability tests alone would not allow the detection of insufficient clinical knowledge. A test of attainment (e.g. the CPST) appears particularly important in this setting. Secondly, candidate perceptions of fairness are not favourable towards generic cognitive ability tests with regard to job relevance. This finding supports research on applicant reactions in other occupations in which cognitive ability tests were not positively received by candidates[26,27] and were perceived to lack relevance to any given job role.[28] Such negative perceptions among candidates can result in undesirable outcomes, such as the loss of competent candidates from the selection process,[29] which has a subsequent negative effect on the utility of the selection process.[30] Furthermore, extreme reactions may lead to an increased propensity for legal case initiation by candidates.[31] By contrast, the CPST received the most approving feedback from candidates compared with all the other tests (NVMA, CATB and SJT) and its immediate job relevance and fairness were favourably perceived. In this context, rejecting applicants to specialty training on the basis of generalised cognitive ability tests alone may be particularly sensitive

because the non-selection of candidates based on cognitive ability test scores may be at odds with their previous achievement of high academic success that enabled their initial selection into medical school. Our findings may suggest that there is a trade-off between costs and time constraints, positive candidate perceptions and using tests that identify clinical knowledge and clinically relevant attributes. This represents a dilemma in terms of balancing the need to reduce costs and administrative and development time, while ensuring that the most appropriate knowledge, skills and attributes are assessed during selection in a manner that is also acceptable to candidates.

Finally, results showed that the two current selection tests (the CPST and SJT) assess cognitive ability to an extent, but that they also assess other constructs likely to relate more closely to behavioural domains that are important in the selection of future general practitioners, such as empathy, integrity and clinical expertise.[16] There is an added security risk associated with using off-the-shelf cognitive ability tests because they can be accessed directly from the test publishers and are susceptible to coaching effects. By contrast, there is a reduced risk with the CPST and SJT as they both adopt an item bank approach and access is more closely regulated.

This study demonstrates that the CPST and SJT are more effective tests than the NVMA and CATB for the purposes of selection into GP training. In combination, these tests provide improved predictive validity compared with the NVMA and CATB, but they are also perceived as having greater relevance to the target job role. It should be noted that the present paper represents a first step towards demonstrating the predictive validity of the cognitive ability tests as the outcome measure used was SC performance. The ultimate goal in demonstrating good criterion-related validity is to predict performance in postgraduate training; therefore future research might investigate the predictive validity of the NVMA and CATB by exploring subsequent performance in the job role. The methods used currently in the GP selection process have been shown to predict subsequent training performance[14] and supervisor ratings 12 months into training,[24] and these outcome criteria may be useful in future research.

REFERENCES

1 Hunter JE. Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *J Vocat Behav* 1986;**29** (3):340–62.

2 Schmidt FL, Hunter JE. The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychol Bull* 1998;**124** (2):262–74.

3 Robertson IT, Smith M. Personnel selection. *J Occup Organ Psychol* 2001;**74** (4):441–72.

4 Schmidt FL. The role of general cognitive ability and job performance: why there cannot be a debate. *Hum Perform* 2002;**15** (1/2):187–210.

5 Campbell JP. An overview of the army selection and classification project (Project A). *Pers Psychol* 1990;**43** (2):232–9.

6 Martinussen M. Psychological measures as predictors of pilot performance: a meta-analysis. *Int J Aviat Psychol* 1996;**6** (1):1–20.

7 Ferguson E, James D, Madeley L. Factors associated with success in medical school: systematic review of the literature. *BMJ* 2002;**324** (7343):952–7.

8 James D, Yates J, Nicholson S. Comparison of A level and UKCAT performance in students applying to UK medical and dental schools in 2006: cohort study. *BMJ* 2010;**340**:478.

9 Lynch B, MacKenzie R, Dowell J, Cleland J, Prescott G. Does the UKCAT predict Year 1 performance in medical school? *Med Educ* 2009;**43** (12):1203–9.

10 Patterson F, Baron H, Carr V, Plint S, Lane P. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Med Educ* 2009;**43** (1):50–7.

11 Patterson F, Carr V, Zibarras L, Burr B, Berkin L, Plint S *et al.* New machine-marked tests for selection into core medical training: evidence from two validation studies. *Clin Med* 2009;**9** (5):417–20.

12 Patterson F, Zibarras LD, Carr V, Irish B, Gregory S. Evaluating candidate reactions to selection practices

19 using organisational justice theory. *Med Educ* 2011;**xx**:xxx–xx.

13 Zibarras LD, Patterson F. Applicant reactions, perceived fairness and ethnic differences in a live, high-stakes selection context. Division of Occupational Psychology Conference, dd–dd Month 2009, Xxx.

14 Patterson F, Ferguson E, Norfolk T, Lane P. A new selection system to recruit general practice registrars: preliminary findings from a validation study. *BMJ* 2005;**330**:711–4.

15 Raven J, Raven JC, Court JH. *Manual for Raven's Progressive Matrices and Vocabulary Scales. The Advanced Progressive Matrices.* Oxford: Oxford Psychologists Press 1998.

16 Patterson F, Ferguson E, Lane P, Farrell K, Martlew J, Wells A. A competency model for general practice: implications for selection, training, and development. *Br J Gen Pract* 2000;**50**:188–93.

17 Hopton T, MacIver R, Saville P, Kurz R. *Analysis Aptitude Range Handbook.* St Helier: Saville Consulting Group 2010.

18 Cattell RB. *Intelligence: Its Structure, Growth, and Action.* New York, NY: Elsevier Science Publishing 1987; xxx–xx.

19 Lievens F, Peeters H, Schollaert E. Situational judgement tests: a review of recent research. *Pers Rev* 2008;**37** (4):426–41.

20 Christian MS, Edwards BD, Bradley JC. Situational judgement tests: constructs assessed and a meta-analysis of their criterion-related validities. *Pers Psychol* 2010;**63**:83–117.

21 McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. Use of situational judgement tests to predict job performance: a clarification of the literature. *J Appl Psychol* 2001;**86** (4):730–40.

22 Weekley JA, Jones C. Video-based situational testing. *Pers Psychol* 1997;**50** (1):25–49.

23 Motowidlo SJ, Beier ME. Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgement test. *J Appl Psychol* 2010;**95** (2):321–33.

24 Lievens F, Patterson F. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced level high-stakes selection. *J Appl Psychol* 2011;**xx**:xxx–xx.

25 Bauer TN, Truxillo DM, Sanchez RJ, Craig JM, Ferrara P, Campion MA. Applicant reactions to selection: development of the selection procedural justice scale (SPJS). *Pers Psychol* 2001;**54** (2):387–419.

26 Anderson N, Witvliet C. Fairness reactions to personnel selection methods: an international comparison between the Netherlands, the United States, France, Spain, Portugal, and Singapore. *Int J Select Assess* 2008;**16** (1):1–13.

27 Nikolaou I, Judge TA. Fairness reactions to personnel selection techniques in Greece: the role of core self-evaluations. *Int J Select Assess* 2007;**15** (2): 206–19.

28 Smither JW, Reilly RR, Millsap RE, Pearlman K, Stoffey RW. Applicant reactions to selection procedures. *Pers Psychol* 1993;**46**:49–76.

29 Schmit MJ, Ryan AM. Applicant withdrawal: the role of test-taking attitudes and racial differences. *Pers Psychol* 1997;**50** (4):855–76.

30 Murphy KR. When your top choice turns you down: effect of rejected offers on the utility of selection tests. *Psychol Bull* 1986;**99**:133–8.

31 Anderson N. Perceived job discrimination (PJD): toward a model of applicant propensity to case initiation in selection. *Int J Select Assess* 2011;**19**: 229–44.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Table S1. Description of the tests and outcome measures used in the study, with examples.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than for missing material) should be directed to the corresponding author for the article.

Author Query Form

Journal: MEDU

Article: 4195

| Query reference	Query	Remarks
1	AUTHOR: Check affiliation linking numbers	
2	AUTHOR: Check affiliation linking numbers Author: check affiliation linking numbers	
3	AUTHOR: Check affiliation linking numbers	
4	AUTHOR: Check affiliation linking numbers	
5	AUTHOR: Check affiliation linking numbers	
6	AUTHOR: Check affiliation linking numbers	
7	AUTHOR: Insert name of department	
8	AUTHOR: Insert city	
9	AUTHOR: Please check the correspondence.	
10	AUTHOR: Check this citation of Table S1 is correct	
11	AUTHOR: Please give software release number	
12	AUTHOR: Please give name of department, institution for PC; include city, country if the institution has not been cited in the author affiliations	
13	AUTHOR: Please give include city, country for Pearson Education	
14	AUTHOR: Please give include city, country for Saville Consulting	
15	AUTHOR: check funding statement	
16	AUTHOR: Insert full name of approving committee	
17	AUTHOR: Please ensure this page range is complete	
18	AUTHOR: If there are fewer than 11 authors for this Reference, please supply all of their names. If there are 11 or more authors, please supply the first 3 author names then *et al.*	
19	AUTHOR: Check year	
20	AUTHOR: Insert volume number	

21	AUTHOR: Insert page range detail	
22	AUTHOR: Insert dates of conference	
23	AUTHOR: Insert city of conference; include state/province abbreviation if in USA, Canada, Australia or Brazil	
24	AUTHOR: Insert page range detail	
25	AUTHOR: Check year	
26	AUTHOR: Insert volume number	
27	AUTHOR: Insert page range detail	
28	AUTHOR: Please check this description is correct	
29	Author: Please check that this is correct or delete all mention of Table S1 and Supporting Information	
30	AUTHOR: check this definition	
31	AUTHOR: Please provide significance for the bolded values in Table 4.	
32	AUTHOR: Figure 1 has been saved at a low resolution of 104 dpi. Please resupply at 600 dpi. Check required artwork specifications at http://authorservices.wiley.com/submit_illust.asp?site=1	

USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

Required software to e-Annotate PDFs: <u>Adobe Acrobat Professional</u> or <u>Adobe Reader</u> (version 8.0 or above). (Note that this document uses screenshots from <u>Adobe Reader X</u>)
The latest version of Acrobat Reader can be downloaded for free at: <u>http://get.adobe.com/reader/</u>

Once you have Acrobat Reader open on your computer, click on the Comment tab at the right of the toolbar:

												Tools	Comment	Share
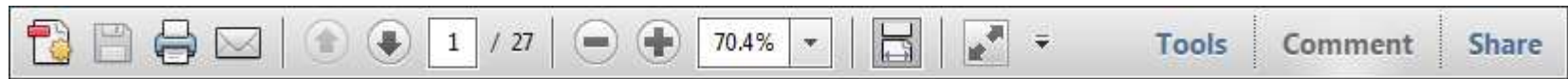
This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the Annotations section, pictured opposite. We've picked out some of these tools below:
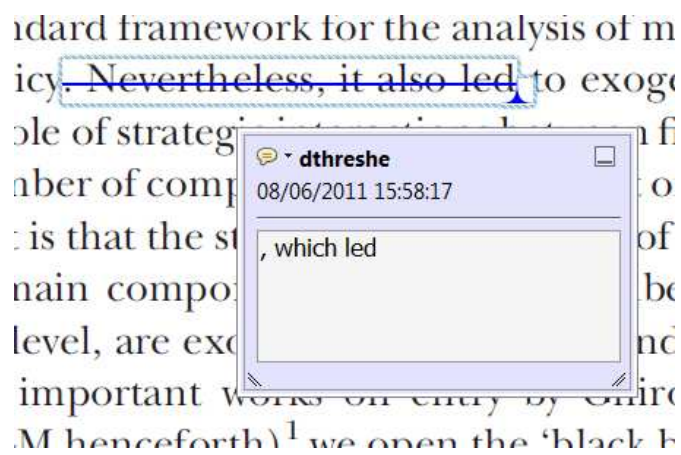
▾ Annotations

---
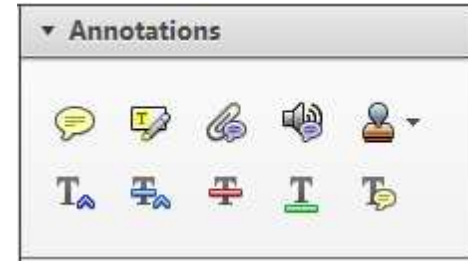
1. Replace (Ins) Tool – for replacing text.

Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it

- Highlight a word or sentence.
- Click on the Replace (Ins) icon in the Annotations section.
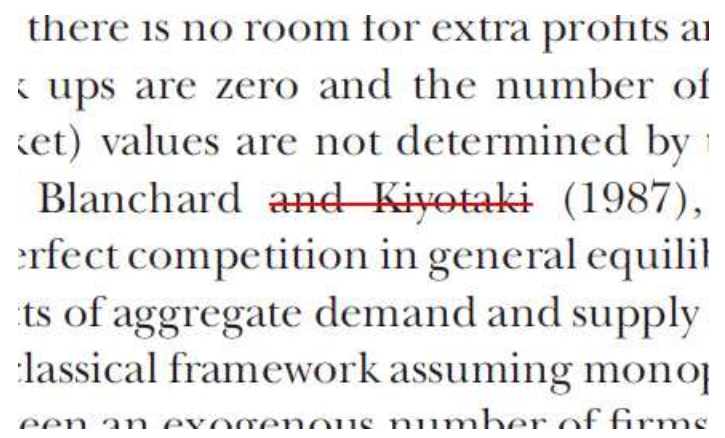- Type the replacement text into the blue box that appears.

dard framework for the analysis of m
icy. Nevertheless, it also led to exoge
ole of strategic interaction between fi
nber of comp
: is that the st
nain compo
level, are exc
important works on entry by Ghird
.M henceforth)[1] we open the 'black l

> dthreshe
> 08/06/2011 15:58:17
> , which led

---

2. Strikethrough (Del) Tool – for deleting text.

Strikes a red line through text that is to be deleted.

How to use it

- Highlight a word or sentence.
- Click on the Strikethrough (Del) icon in the Annotations section.

there is no room for extra profits ar
: ups are zero and the number of
:ket) values are not determined by :
Blanchard and Kiyotaki (1987),
rfect competition in general equilil
ts of aggregate demand and supply
:lassical framework assuming monoj
een an exogenous number of firms

---

3. Add note to text Tool – for highlighting a section to be changed to bold or italic.

Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the Add note to text icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.
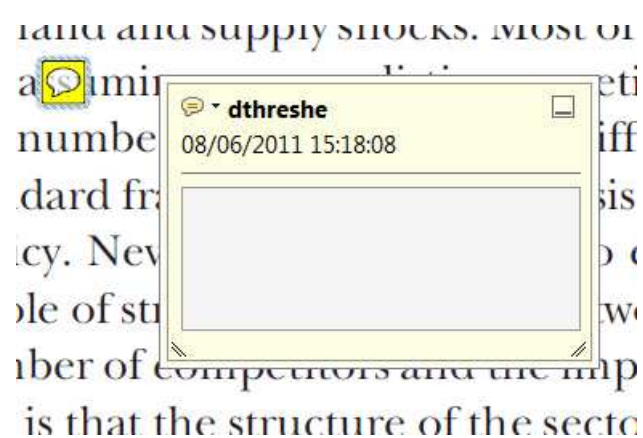
namic responses of mark ups
ent with the **VAR** evidence

sation
y Ma
and
on n
to a
stent also with the demand-

> dthreshe
> 08/06/2011 15:31:38

---

4. Add sticky note Tool – for making notes at specific points in the text.

Marks a point in the proof where a comment needs to be highlighted.

How to use it

- Click on the Add sticky note icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.

land and supply shocks. Most of
a :amin
numbe
dard fra
:cy. New
ole of st
nber of competitors and the imp
: is that the structure of the secto

> dthreshe
> 08/06/2011 15:18:08

# USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

---

5. Attach File Tool – for inserting large amounts of text or replacement figures.

Inserts an icon linking to the attached file in the appropriate pace in the text.

How to use it

- Click on the Attach File icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.
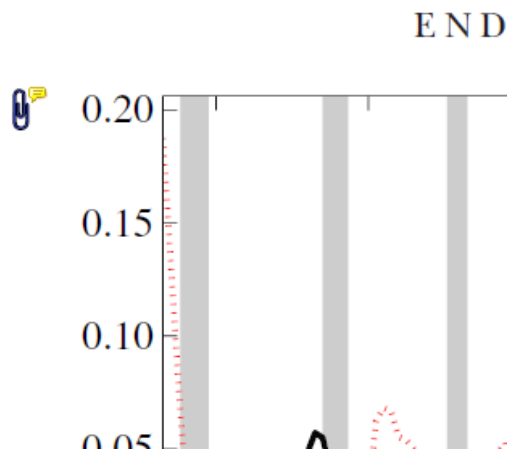
6. Add stamp Tool – for approving a proof if no corrections are required.

Inserts a selected stamp onto an appropriate place in the proof.

How to use it

- Click on the Add stamp icon in the Annotations section.
- Select the stamp you want to use. (The Approved stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).


---

7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

How to use it

- Click on one of the shapes in the Drawing Markups section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.
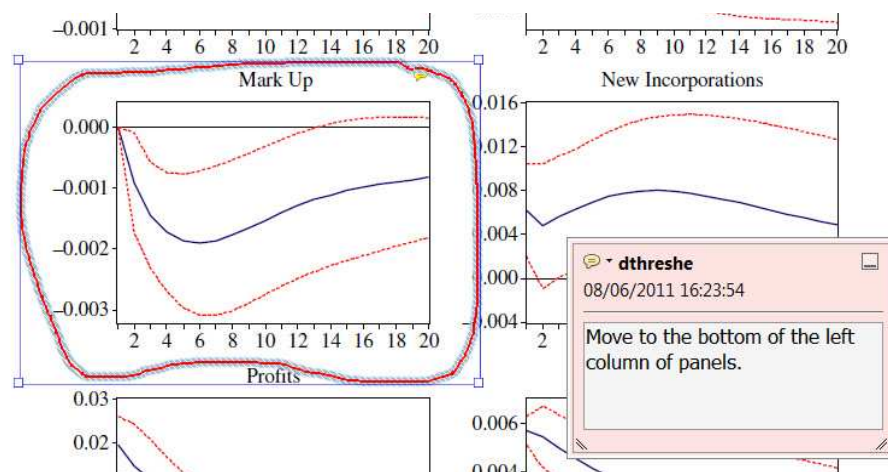
For further information on how to annotate proofs, click on the Help menu to reveal a list of further options: