medical education in review

Evaluations of situational judgement tests to assess non-academic attributes in selection

Fiona Patterson,[1,2] Victoria Ashworth,[2] Lara Zibarras,[3] Philippa Coan,[2] Maire Kerrin[2] & Paul O'Neill[4]

OBJECTIVES This paper presents a systematic review of the emerging international research evidence for the use of situational judgement tests (SJTs) for testing important non-academic attributes (such as empathy, integrity and resilience) in selection processes.

METHODS Several databases (e.g. MEDLINE, PsycINFO, Web of Science) were searched to retrieve empirical studies relating to SJTs published between 1990 and 2010. Personal contact with experts in the field was made to identify any unpublished research or work in progress to obtain the most current material. Finally, reference lists were checked to access other relevant journal articles and further research. All research studies were required to meet specific inclusion criteria selected by two independent reviewers.

RESULTS Over 1000 citations were identified during the initial literature search; following the review of abstracts, full-text copies of 76 articles were retrieved and evaluated. A total of 39 articles that adequately met the inclusion criteria were included in the final review. The research evidence shows that, compared with personality and IQ tests, SJTs have good levels of reliability, predictive validity and incremental validity for testing a range of professional attributes, such as empathy and integrity.

CONCLUSIONS SJTs can be designed to test a broad range of non-academic constructs depending on the selection context. As a relatively low-fidelity assessment, SJTs are a cost-efficient methodology compared with high-fidelity assessments of non-academic attributes, such as those used in objective structured clinical examinations. In general, SJTs are found to demonstrate less adverse impact than IQ tests and are positively received by candidates. Further research is required to explore theoretical developments and the underlying construct validity of SJTs.

Medical Education 2012: 46: 850–868
doi:10.1111/j.1365-2923.2012.04336.x

Discuss ideas arising from this article at
www.mededuc.com 'discuss'

[1]Department of Psychology, School of the Biological Sciences, University of Cambridge, Cambridge, UK
[2]Work Psychology Group, Nottingham, UK
[3]Department of Organisational Psychology, School of Social Sciences, City University, London, UK
[4]Department of Elderly Care, School of Medicine, University of Manchester, Manchester, UK

Correspondence: Professor Fiona Patterson, Department of Psychology, Downing Street, Cambridge, CB2 3EB, UK. Tel: 00 44 940 7978; E-mail: fcp27@cam.ac.uk

© Blackwell Publishing Ltd 2012.

Evaluations of situational judgement tests to assess non-academic attributes in selection

OBJECTIVES

This paper presents a systematic review of the emerging international research evidence for the use of situational judgement tests (SJTs) for testing important non-academic attributes (such as empathy, integrity and resilience) in selection processes.

METHODS

Several databases (e.g. MEDLINE, PsycINFO, Web of Science) were searched to retrieve empirical studies relating to SJTs published between 1990 and 2010. Personal contact with experts in the field

was made to identify any unpublished research or work in progress to obtain the most current material. Finally, reference lists were checked to access other relevant journal articles and further research. All research studies were required to meet specific inclusion criteria selected by two independent reviewers.

RESULTS

Over 1000 citations were identified during the initial literature search; following the review of abstracts, full-text copies of 76 articles were retrieved and evaluated. A total of 39 articles that adequately met the inclusion criteria were included in the final review. The research evidence shows that, compared with personality and IQ tests, SJTs have good levels of reliability, predictive validity and incremental validity for testing a range of professional attributes, such as empathy and integrity.

CONCLUSIONS

SJTs can be designed to test a broad range of non-academic constructs depending on the selection context. As a relatively low-fidelity assessment, SJTs are a cost efficient methodology compared with high-fidelity assessments of non-academic attributes, such as those used in objective structured clinical examinations. In general, SJTs are found to demonstrate less adverse impact than IQ tests and are positively received by candidates. Further research is required to explore theoretical developments and the underlying construct validity of SJTs.

INTRODUCTION

Research and practice show that an array of non-academic professional attributes, such as integrity, empathy and team awareness, are critically important predictors of job performance and training outcomes. Until recently, international selection practices tended to focus on assessing academic ability alone.1,2 A key challenge for recruiters concerns how best to assess a broad range of non-academic attributes reliably, as, for example, large-scale interviewing can be costly and the use of personality tests is problematic because there is limited evidence to support their predictive validity for selection purposes in high-stakes settings.1,3 In their international review of selection practices, Prideaux et al.1 ask whether the situational judgement test (SJT) may be a valid and reliable method for assessing a broad range of non-academic attributes in high-volume selection.4 Although SJTs are currently used for selection into many other professions, research evidence for their use in selection in medicine is only just emerging. This review specifically examines the extent to which an SJT (as a comparatively low-fidelity assessment) can truly measure non-academic attributes that are important in clinical practice, such as integrity, empathy and resilience. Furthermore, it investigates whether SJTs, which are often presented in a written examination format, add predictive validity over and above personality tests, tests of IQ and behavioural interviews.

In this paper, we present a systematic review of SJTs as a selection method in different occupations and discuss the specific implications for their potential relevance for selection into medical education and training compared with other more established methods.

Although the SJT method has been in existence for several decades, in the past 10 years the use of these tests has become increasingly popular in large-scale selection across all occupational groups internationally. The SJT is a measurement method rather than a single style of test presentation; it is designed to assess an applicant's judgement in situations encountered in the workplace in a manner that targets professional non-academic attributes rather than clinical knowledge.5 In an SJT, applicants are presented with written or video-based depictions of hypothetical scenarios and asked to identify an appropriate response from a list of alternatives. Table 1 shows example items in SJTs

used in selection for medical training. A variety of response formats can be used and are typically classified into one of two format types according to whether they are knowledge-based (i.e. 'What is the best option?') or refer to behavioural tendency (i.e. 'What would you be most likely to do?'). Response alternatives can be presented in either a written (low-fidelity) or a video-based (medium-fidelity) format.5,6 Situational judgement tests are typically scored by comparing applicants' responses with a predetermined scoring key agreed by subject matter experts. As a selection tool, the SJT is growing in popularity because it has useful levels of face, content and predictive validity6,7 and scenarios used in SJTs are typically derived from job analysis studies.8 Broadly, SJTs are based on behavioural consistency theory, which asserts that past behaviour is the best predictor of future behaviour.9

SJTs are an established method of selection for use in contexts of high-volume selection in many occupational groups5 and research shows that they can be used reliably to select for a range of non-academic or professional attributes.10,11 Although other selection methods, such as panel interviews or personality questionnaires,12 may also be used to select for these attributes, SJTs may offer significant advantages over these methods. For example, panel interviews are often criticised for their potential for bias or because they lack standardisation,13 and personality tests offer lower face validity and are less acceptable to candidates as a selection tool.14 By contrast, SJTs offer a standardised method of objectively assessing a broad range of non-academic attributes in large numbers of applicants, and demonstrate face validity to candidates because the scenarios used in SJTs are based on job-relevant situations.6 There is emerging evidence to show that SJTs represent a useful method of evaluating a range of professional attributes for selection into medicine, for both admission to medical school and entry into postgraduate training.8,15,16 For example, an SJT measuring empathy, integrity and resilience is used to select candidates applying for training in UK general practice8,17 and an SJT is used to measure applicants' interpersonal awareness in medical school admissions in Belgium.18 Not only might the SJT offer an objective way of reliably assessing these non-academic attributes, but because the SJT is a machine-markable selection method, it may also offer significant cost savings over other methods that are hand-scored by trained assessors, such as in selection centre exercises.16

To date, there has been no systematic review of the current research literature commenting on the potential utility of SJTs in medical education and training. Hence, the present review complements previous reviews of SJTs5,19 in several ways. Firstly, we focus on the relevance of SJTs for use in high-stakes selection as in medical education and training. Secondly, we address not only the psychometric properties of SJTs, but also aspects more pertinent to selection into medicine, such as candidate perceptions, cost-effectiveness in high-volume selection and susceptibility to coaching. As the evidence base for SJTs in medical education is currently relatively sparse, we draw on research and literature from all occupational groups to explore implications for future research.

METHODS

Data sources

Five databases (MEDLINE, PsycINFO, Web of Science, BIOME and BioMed Central) and seven peer reviewed journals that publish selection research (Journal of Applied Psychology, International Journal of Selection and Assessment, Personnel Psychology, Medical Education, International Journal

of Human Resource Management, Journal of Occupational and Organisational Psychology and Medical Teacher) were used to conduct a computer-based search of literature spanning

Table 1 Example items on situational judgement tests (SJTs) for medical selection showing different response formats

Table 1 Example items on situational judgement tests (SJTs) for medical selection showing different response formats

SJT item using a multiple-response format	SJT item using a ranking-response format	SJT item using a best-response format
You are looking after Mrs Sandra Jones, who is being investigated in hospital. You are asked by her family not to inform Mrs Jones if the results confirm cancer *Choose the THREE most appropriate actions to take in this situation* (a) Ignore the family's wishes (b) Agree not to tell Mrs Jones (c) Explain to the family that it is Mrs Jones' decision (d) Ask Mrs Jones whether she wishes to know the test results (e) Ask Mrs Jones whether she wishes you to inform the family (f) Inform Mrs Jones that her family does not wish her to have the results (g) Give the results to the family first (h) Give the results to the next of kin first	You are reviewing a routine drug chart for a patient with rheumatoid arthritis during an overnight shift. You notice that your consultant has inappropriately prescribed methotrexate 7.5 mg daily instead of weekly *Rank in order the following actions in response to this situation (1 = most appropriate; 5 = least appropriate)* (a) Ask the nurses if the consultant has made any other drug errors recently (b) Correct the prescription to 7.5 mg weekly (c) Leave the prescription unchanged until the consultant ward round the following morning (d) Phone the consultant at home to ask about changing the prescription (e) Inform the patient of the error	Patient: So, this physiotherapy is really going to help me? Doctor: Absolutely, even though in the first days it might still be painful Patient: Yes, I suppose it will take a while before it starts working Doctor: That is why I am going to prescribe a painkiller. You should take three painkillers per day Patient: Do I really have to take them? I have already tried a few things. First, they didn't help me. And second, I'm actually opposed to taking any medication. I'd rather not take them. They are not good for my health *What is the best way for you (as a doctor) to react to this patient's refusal to take the prescribed medication?* (a) Ask her if she knows of something else to relieve the pain (b) Give her the scientific evidence as to why painkillers will help (c) Agree to her not taking them now, but stress the importance of the physiotherapy (d) Tell her that, in her own interests, she will have to start changing her attitude

To retrieve empirical studies relating to SJTs, search terms included the following keywords: situational judgement test; select*; valid*; reliab*, and fairness. In addition, reference lists from papers were manually checked to seek other relevant journal articles and experts in the field were contacted for unpublished studies and works in progress.

Selection of studies and inclusion criteria

A researcher independently reviewed the abstracts identified in the initial search (n > 1000) to investigate whether the inclusion criteria were met in each. These inclusion criteria required papers to be peer reviewed articles (therefore general opinion pieces and commentaries were excluded), to be studies using empirical data or reviews or meta-analyses focusing exclusively on SJTs, and to use

occupational samples. This process of examination to establish whether or not articles fulfilled the inclusion criteria identified 76 articles that did so. Full-text versions of these 76 articles were retrieved and reviewed independently by two researchers.

Data abstraction

During the full-text review, data extraction sheets were used to record all relevant information. The data extraction sheets comprised a table recording the following: method of identification (e.g. database); author(s); title; year; volume; pages; type of article (e.g. empirical, literature review, meta-analysis); research aim; study design; analysis; key focus of the paper (e.g. reliability, validity, fairness), and main findings and results. If an article was not deemed to be relevant following this review, a note was made to this effect.

RESULTS

Over 1000 citations were retrieved during the initial search. Following a review of these citations, the full texts of 76 papers were retrieved and independently reviewed by two researchers. The same inclusion criteria were applied to the full-text papers to identify those peer-reviewed articles (reviews, empirical studies or meta-analyses) that focused on the use of SJTs in occupational settings or samples. After the inclusion criteria had been applied to the full-text papers, a total of 39 independent articles were retained for this final review. Note that some articles were discarded because they represented duplications. As only a small number of the articles retrieved were set in a medical context (nine of the 39 studies), any formal meta-analysis was ruled out. The following results are structured under the headings of evidence regarding: (i) reliability; (ii) validity (criterion, incremental and construct validity); (iii) fairness; (iv) candidate reactions; (v) cost-effectiveness, and (vi) susceptibility to coaching. A brief summary of research findings along with details of the context and sampling of the study reported in each article included in this review is presented in Appendix 1.

How reliable are SJTs?

There is an inherent difficulty in accurately estimating the reliability of an SJT because SJTs typically measure multiple constructs. SJTs are often described as being 'construct heterogeneous' at the item level,19 because one item, for example, may target several performance dimensions. In general, researchers tend to report good levels of internal consistency for SJTs. In McDaniel et al.'s meta-analysis,20 39 studies of different SJTs found internal consistency (Cronbach's alpha) coefficients ranging from a = 0.43 to a = 0.94. However, SJT length moderates reliability and SJTs that include greater numbers of items are more reliable, which is consistent with evidence21 that internal reliability is influenced by scale length. The type of response instructions used also appears to influence reliability. For example, Ployhart and Ehrhart22 found that the highest internal consistency (a = 0.73) was obtained when candidates were asked to rate the effectiveness of each response; conversely, when candidates were asked to choose the most effective response, the lowest internal consistency (a = 0.24) was found. Note that in this research, each response option type yielded the same number of data points because a score per item was used and thus reliability was not simply a matter of length of test. In selection for UK postgraduate training in general practice, Patterson et al.8 found internal consistencies (Cronbach's alpha) to range from a = 0.80 to a = 0.83 for an SJT piloted for selection into postgraduate training (n = 2552).

As SJTs are multifaceted in nature, some researchers contend that internal consistency is not the best way to measure reliability because internal consistency is more appropriate for uni-dimensional tests.5,23 Researchers argue that test–retest or parallel forms reliability may be a more accurate approach to evaluating the reliability of any given SJT. Empirical research shows a range of test–

retest and parallel form reliabilities reported for different SJTs. For example, Ployhart et al.24 found a test–retest reliability of r = 0.84, whereas Ployhart and Ehrhart22 report test–retest reliabilities ranging from as low as r = 0.20 to as high as r = 0.92 depending on the type of response instructions used, with higher reliabilities when candidates were asked to rate the likelihood that they would perform each option. Chan and Schmitt25 report parallel form reliability of r = 0.76, and Lievens et al.26 report r = 0.66 for two alternative forms of an SJT developed for use in medical school admissions. Overall, research generally shows moderate to good levels of reliability for SJTs, whichever method is used.

How valid are SJTs?

Criterion-related validity

Several empirical studies have shown SJTs (across a range of occupations) to be effective measures of subsequent job performance and training criteria. For example, Chan and Schmitt25 showed SJTs to have substantial validity in predicting task performance (core technical proficiency), contextual performance (job dedication, interpersonal facilitation) and overall job performance. In validating an SJT for measuring personal initiative, researchers27 found significant correlations with supervisor ratings of overall performance (r = 0.37) and personal initiative specifically (r = 0.48). Borman et al.28 found a validity coefficient of r = 0.22 with supervisory ratings of performance in the military. Other researchers have used objective criterion measures in, for example, sales performance,10 in which significant correlations were obtained between performance on an SJT and sales performance ranking (r = 0.63). In medicine, an SJT piloted in selection for postgraduate training showed substantial criterion-related validity with a correlation of r = 0.52 between applicant scores on the SJT and subsequent performance ratings in a multiple-mini interview.15

McDaniel et al.'s meta-analysis,20 based on 102 validity coefficients across many different occupational groups, estimated that SJTs have an overall criterion-related validity of r = 0.34. This is comparable with assessment centres,29 structured interviews13 and biographical data, such as previous qualification and experience.30 When corrected for range restriction and criterion unreliability, the validity of SJTs increased to r = 0.56, which is similar to that of cognitive ability tests.12 However, McDaniel et al.'s meta-analysis20 shows substantial variability of coefficients, with SJTs based on a job analysis generally showing substantially higher validity than those that are not (r = 0.38 versus r = 0.29). More recently, McDaniel et al.31 reported the overall uncorrected validity coefficient of SJTs across 118 coefficients to be r = 0.26 (n = 24 756), regardless of instruction type.

Incremental validity

It is also theoretically and practically important to examine an SJT's predictive validity over and above other established selection methods; this is referred to as incremental validity and is usually analysed using hierarchical regression techniques. A number of different studies have examined whether SJTs add significant variance to the prediction of job performance over other selection methods, such as cognitive ability, job knowledge and personality.32 For example, Lievens et al.18 examined the incremental validity of an SJT that aimed to measure non-academic attributes (interpersonal awareness and communication) over traditional cognitively oriented measures for medical school admissions in Belgium. The SJT showed significant incremental validity over the cognitively oriented predictors for curricula components that had a substantial interpersonal skill component (such as communication skills), but not for curricula with a minimal interpersonal skill

component. In addition, the validity of the SJT increased through the four academic years, which is consistent with research showing that non-academic predictors become more important when the criterion data are gathered at a later date, such as during clinical practice.33,34 Lievens et al.18 concluded that SJTs can be a useful and valid complement to traditional student admission tests, even in an operational high-stakes context.

Using hierarchical regression, O'Connell et al.32 found that SJTs had incremental validity over cognitive and personality measures in predicting task performance (adding 3% and 4% of variance, respectively), but had incremental validity over cognitive ability only in predicting contextual performance (4%). Patterson et al.8 compared the validity of three shortlisting methodologies for selection into postgraduate general practice training in a context in which an SJT was piloted alongside an application form and a clinical problem-solving test. All three tests predicted subsequent performance in job simulations at a selection centre, but the SJT was the best single predictor, showing substantial incremental validity over the application form scores (predicting an additional 17% of the variance in selection outcomes) and the clinical problem-solving test (predicting an additional 14% of the variance).

The results of McDaniel et al.'s meta-analysis31 examining the incremental validity of SJTs show that SJTs provide significant incremental validity over measures of cognitive ability (3–5%) and over measures of personality (6–7%). Thus, SJTs provide significant incremental validity over cognitive ability and personality measures combined. Patterson et al.15 showed that an SJT offered significant incremental validity (17%) over other shortlisting methods for selection into core medical training. In a recent longitudinal predictive validation study tracking trainees in UK general practice, Lievens and Patterson16 found that the SJT (compared with a knowledge test and a high-fidelity selection centre) was the best independent predictor of subsequent job performance after 12 months in training, measured using supervisor ratings.

Construct-related validity

Empirical research suggests that SJTs may be related to a variety of constructs. This is perhaps unsurprising as SJT items may relate to different situations and include different content domains, and a candidate's response judgement may relate to a combination of his or her personality, ability and experience.5 As such, empirical research shows SJTs to be correlated with personality,35 cognitive ability19 and job knowledge.36 For example, Mullins and Schmitt35 found that an SJT was most strongly correlated with both of the personality factors Conscientiousness (r = 0.26) and Agreeableness (r = 0.22), and Smith and McDaniel37 found their SJT correlated with measures of Conscientiousness (r = 0.32) and Emotional Stability (r = 0.22) and found no relationship between the SJT and Agreeableness. In a meta-analysis, McDaniel and Nguyen23 examined correlations between SJT scores and personality characteristics. They found aggregated correlations with SJT scores of r = 0.25 for Agreeableness, r = 0.26 for Conscientiousness, r = 0.31 for Emotional Stability, r = 0.06 for Extroversion, and r = 0.09 for Openness. Overall, there appears to be some evidence to suggest that SJTs may be correlated with important aspects of personality, but exact relationships are unclear.

McDaniel et al.20 showed that SJTs also correlate with cognitive ability (r = 0.46); similarly, these authors' meta-analysis showed some variability around this finding. An examination of relationship moderators indicated only modest success in explaining this finding; for example, SJTs based on a job analysis were more highly correlated with cognitive ability (r = 0.50 versus r = 0.38), and SJTs with more detailed questions were more highly related to cognitive ability (r = 0.56 versus r = 0.47). However, research suggests that SJT scores are uncorrelated with cognitive ability.11,25 Clevenger

et al.11 suggest that variability in estimates of the correlation between cognitive ability and SJTs may reflect the nature of the situations presented. Those tasks that are more cognitively oriented (e.g. planning, organising, analytical problem solving) correlate more highly with cognitive ability than those that refer to interpersonal attributes (e.g. empathy), in which personality may play a greater role. There is also evidence to suggest that SJT format has an influence on construct validity. For example, McDaniel et al.31 showed that SJTs with knowledge-oriented response instructions (i.e. 'What should you do?') had higher correlations with cognitive ability tests (r = 0.35) than those with behavioural tendency instructions (i.e. 'What would you do?') (r = 0.19). Conversely, behavioural tendency SJTs had higher correlations with Agreeableness (r = 0.37), Conscientiousness (r = 0.34) and Emotional Stability (r = 0.35) than those with knowledge instructions (r = 0.19, r = 0.24 and r = 0.12, respectively). Thus, it could be inferred that SJTs with knowledge-based instructions may be considered measures of idealised performance, whereas those with behavioural tendency instructions may be typical performance measures.5 In a postgraduate selection context, Koczwara et al.38 found an SJT to correlate significantly with tests of cognitive ability, which suggests that they measure overlapping constructs.

There has been significant ongoing debate regarding the association between SJTs and job knowledge, in which some researchers argue that SJTs may represent merely tests of job knowledge.39 Other researchers propose that as 'experience' is a multifaceted construct, different operationalisations may result in differential findings regarding the relationship between SJTs and job experience (e.g. general versus highly job-specific experience).40,41 Weekley and Jones36 report a correlation of r = 0.20 between an SJT and job experience and, in a later study, the same authors42 report a correlation of r = 0.23 between SJT scores and a broad measure of general work experience. These authors argue that some SJTs may be more strongly related to general experience measures than to highly job-specific experience measures.42

How fair are SJTs?

The issue of adverse impact and making assessment fair and equal in medical education is fiercely debated.43 In dealing with issues of fairness, selection researchers generally examine adverse impact: that is, the extent to which particular groups systematically perform poorly compared with other groups on a particular selection method. Currently, the research relating to SJTs has generally focused on ethnicity and gender effects.

In relation to ethnicity, differences in mean scores between different ethnic groups are much smaller for SJTs than for tests of cognitive ability.44,45 This is encouraging because the issue of adverse impact of assessments in medical education is problematic46 and adverse impact in cognitive ability tests is well documented in the selection literature, in which findings suggest there is a difference of one standard deviation (SD) in mean test scores in favour of White applicants.47,48 In a meta-analysis of SJT research across all occupational groups, Nguyen et al.49 found the mean difference in SJT scores between Black and White groups was 0.38 SD in favour of White candidates; however, subgroup differences were reduced substantially if SJTs mostly measured non-academic aspects of job performance. Results suggest that behavioural tendency instructions have lower adverse impact than knowledge-based formats.7 Chan and Schmitt50 showed that video-based SJTs had less adverse impact than written SJTs because they were less cognitively loaded. The authors demonstrated this by holding the test content constant in both the written and video-based SJTs; additionally, a standard reading comprehension test was included that allowed a direct test of the notion that the cognitive loading attributable to comprehension accounted for the change in subgroup differences across methods of assessment.50 In a meta-analysis, Whetzel et al.7 found mean ethnic differences in SJT performance. Their findings revealed that, on average, White test-

takers performed better on SJTs than Black (d = 0.38), Hispanic (d = 0.24) and Asian (d = 0.29) candidates.7 O'Connell et al.32 also found similar differences in SJT scores between Black and White groups.

In relation to sex, generally findings indicate that females appear to score slightly higher on SJTs than males, which is consistent with outcomes in other forms of assessment,43 such as assessment centres.51 Nguyen et al.49 found that females score slightly better than males, with a difference in mean scores of 0.1 SD in favour of females. O'Connell et al.32 found a sex difference of 0.27 in favour of females, and Whetzel et al.7 also reported that female candidates scored slightly better than male candidates (d = 0.11).

What are candidate reactions to SJTs?

Applicants demonstrate a preference for job-related selection methods52,53 and research shows that work sample selection methods receive positive ratings by applicants.54 Therefore, it is perhaps not surprising that research into applicant perceptions of SJTs shows SJTs to be perceived positively by candidates.55 Some research shows that video-based SJT formats receive more positive ratings than written formats,50 whereas other research indicates that, in a high-stakes context, applicant reactions to both types of SJT format were relatively equally positive.54 Kanning and Kuhne56 investigated applicants' perceptions of different types of SJT items and reported findings that suggested that interactive SJTs that used videos in both the stimulus and the response elements received the most positive ratings from applicants. In a postgraduate medical selection context, Patterson et al.57 found that applicants' reactions to various selection methods (an SJT, a knowledge test and a selection centre) were most positive for the higher-fidelity simulations in the selection centre. Applicant reactions to the SJT (which required broader judgements based on elements of procedural knowledge relating to both clinical aspects and interpersonal relationships with colleagues, patients and supervisors) were initially more mixed than their responses to the knowledge test (which required judgements based on declarative knowledge). However, over the course of three successive annual recruitment rounds, applicant reactions to the SJT became more positive; this may in part have reflected the fact that more detailed information about SJTs was provided in advance of testing.

How cost-effective are SJTs?

Researchers suggest that SJTs are relatively easy to administer and score5 and have been shown to be cost-effective measures of job performance in a range of different organisations.8,10 Indeed, an important advantage of SJTs is that they can be delivered to applicants in large-group sessions, thus diminishing the costs associated with test administration.5 Recent research8 has shown how the use of a machine-marked SJT as a shortlisting method in postgraduate medical selection (in which the individual marking of thousands of questions is not an efficient or effective use of selectors' time and resources) can deliver significant cost savings in comparison with an application form.

Are SJTs susceptible to coaching and practice effects?

When selection procedures become popular, candidates may attend commercial coaching programmes to learn how to enhance their test scores, as they often do in medicine. This raises the question of whether SJT performance can be improved through coaching. This issue is yet to be extensively explored in the literature. Cullen et al.58 examined the coaching susceptibility of two SJTs (the College Student Questionnaire [CSQ] and the Situational Judgement Inventory [SJI]). They developed a coaching course for these two SJTs that identified test-taking strategies. Findings indicate that, with coaching, scores increase for the CSQ but decrease for the SJI (compared with

scores in controls). In order to elucidate these findings, the authors administered a follow-on survey asking participants about the level of effort required to understand the test-taking strategies and found that strategies for the SJI were considered more difficult.58 The authors concluded that SJTs constructed from subject matter expert judgements are less susceptible to coaching effects.58

A similar issue relates to whether candidates can improve scores when they retake an SJT. Lievens et al.18 investigated the effects of retesting for medical schools admissions in an operational context. The authors demonstrated that retesting effects did exist, but were no larger than those for cognitive ability tests.18 In addition, there was no significant change to the operational validity of the SJT.18 One way of reducing the possibility of coaching and practice effects is to use an item banking approach or to create alternative forms of the SJT. It is worth noting that this approach has been used in research and practice in medical selection38 to reduce susceptibility to coaching.

DISCUSSION

There is an emerging consensus59 that the SJT is essentially a measurement method that can be used to measure a variety of constructs and that its primary targets are non-academic attributes. As a profession, medicine has generally tended to focus on assessing academic abilities; however, research4 has demonstrated the importance of testing non-academic attributes such as interpersonal skills, empathy, integrity and teamwork in order to predict successful job performance in medical careers. The review of the literature shows that SJTs tend to be robust in terms of reliability and validity and therefore present a useful way of measuring these attributes in large-volume selection contexts. However, research also suggests that high-fidelity assessments (such as patient simulation exercises in assessment centres) can provide predictive validity over SJTs (which are low-fidelity assessments) for interpersonally oriented performance domains.16 For example, Lievens and Patterson16 showed that an assessment centre provided incremental validity over an SJT in predicting outcomes in key interpersonal domains such as empathy and communication skills for postgraduate selection. Furthermore, in exploring the optimum combination of selection methods, selection system design should not focus on establishing how much validity a specific selection method adds, but, rather, on considering for what the method in question is valid.

We argue that it is more useful to consider the construct domain underlying a given SJT than to view the SJT as a single measurement tool. By extension, predicting the relative validity (and utility) of any given SJT requires that attention is given to the criterion construct in question. For example, an SJT that focuses on interpersonal performance would not be expected, theoretically, to predict a cognitively loaded criterion measure. As a result, research needs to disentangle the effects of the measurement method from the constructs being targeted17 and more readily address this 'criterion problem'. Practically, several authors have recommended a construct-oriented paradigm for future SJT research which decrees that researchers should report detailed construct information so that it is easier to map and compare the impacts of using different types of stimulus material (e.g. paper and pencil-based, computerised), response modalities (e.g. written, oral) and scoring strategies (subject matter experts, consensus-based scoring) whilst holding the construct constant.

Further research is required to establish greater conceptual clarity regarding SJTs, particularly with reference to issues surrounding construct validity. This requires an improved theoretical framework and one area of research that offers important insights here is that of implicit trait policies (ITPs). Motowidlo and Beier59 define ITPs as an individual's judgement about the relative costs and benefits of expressing certain traits in certain situations. An example is a context in which judging that generally being agreeable in a situation (e.g. towards a patient, a colleague or a supervisor) might be a more successful strategy for dealing with the situation than being disagreeable. In this

way, SJTs measure procedural knowledge about what is effective behaviour in a given situation. Depending on the context, or level in the education and training pathway, this procedural knowledge measured in SJTs comprises specific job knowledge and the candidate's ITPs, which are drawn from general experiences and personality traits.

ITPs are thought to be shaped by experiences in fundamental processes of socialisation, such as that provided by parental modelling during childhood. This may teach the utility of, for example, agreeable expressions (e.g. helping others in need, turning the other cheek, looking after one's neighbours) or disagreeable expressions (e.g. showing selfish preoccupation with one's own interests, holding a grudge, getting even, advancing one's own interests at another person's expense). As students tend to enter medical school in early adulthood, the challenge for educational supervisors is to teach students the utility of effective behaviour in the role of doctor. This process of socialisation is tutored during supervised clinical practice at medical school and refers to effective behaviour by a doctor in any given situation. How much ITPs change after some period of development (e.g. after early adulthood) is unknown. Further research in this area might explore the extent to which differences in the development of ITPs exist among specialties, whereby individuals in postgraduate education and training may learn to make different judgements about what is effective behaviour according to the area, such as paediatrics or surgery, in which they are training and in which the behaviours may differ. Beyond selection, SJTs may represent a helpful springboard to more focused dialogue with and feedback to trainees about how they make decisions on how best to behave towards patients, seniors and colleagues in work-related contexts.

Overall, the present review has four key implications for medical education and training. Firstly, SJTs can be designed to measure a variety of non-academic attributes beyond clinical knowledge, which are especially relevant in medical education, training and practice.8,38 It is important to note that SJTs are not measures of ethical values per se, but, rather, measures of trainees' awareness about what is effective behaviour in work-relevant contexts in important interpersonal domains. Secondly, the research evidence shows SJTs to have good predictive validity, over and above tests of IQ and personality, which is likely to translate into significant gains in the utility of selection systems that incorporate them. Thirdly, as SJTs are scored using predetermined keys, the tests themselves can be systematically scored and are often computer-delivered. This can produce substantial cost savings in high-volume selection as the tests can be machine-marked. Finally, SJTs can target non-academic attributes at various levels in the training pathway, from medical school admissions (in which there is no specific job knowledge) through to senior appointments in specialty selection. Furthermore, although SJTs have been more commonly used as tests of aptitude for selection, the emerging evidence suggests that using feedback from SJTs (e.g. regarding an individual's awareness about effective behaviour at work) may also play a part in the education and training process.

Contributors: all authors made substantial contributions to the study design, intellectual content, data interpretation and analysis, and the writing of the paper. FP and VA conceived of the overall study and analysis strategies, and led the interpretation of results. LZ, PC, MK and PO'N assisted in data collection, analysis and interpretation and in drafting the paper. All authors contributed to the critical revision of the paper and approved the final manuscript for publication.

Acknowledgements: the authors wish to acknowledge the comments provided by Dr Katie Petty-Saphon, Dr Denis Shaugnessey and Siobhan Fitzpatrick of the Medical Schools Council, and the support provided by Dr David Good, Department of Psychology, University of Cambridge.

Funding: Department of Health, England, UK. Conflicts of interest: FP, VA, PC and MK provide advice on selection methodology to the Department of Health through the Work Psychology Group.

Ethical approval: not required.

REFERENCES

1 Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, Patterson F, Powis D, Tekian A, Wilkinson D. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 conference. Med Teach 2011;33 (3):215–23.

2 Eva K, Reiter H, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ 2009;43:767–75.

3 Albanese M, Snow M, Skochelak S, Huggett K, Farrell P. Assessing personal qualities in medical school admissions. Acad Med 2003;78 (3):313–21.

4 Patterson F, Lievens F, Kerrin M, Zibarras L, Koczwara A. Designing selection systems for medicine: implications for the political validity of high-stakes selection methods. Int J Select Assess 2012; in press.

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

6 Christian M, Edwards B, Bradley J. Situational judgement tests: constructs assessed and a meta-analysis of their criterion-related validities. Pers Psychol 2010;63: 83–117.

7 Whetzel D, McDaniel M, Nguyen N. Subgroup differences in situational judgement test performance: a meta-analysis. Hum Perform 2008;21:291–309.

8 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.

9 Motowidlo SJ, Dunnette MD, Carter GW. An alternative selection procedure: the low-fidelity simulation. J Appl Psychol 1990;75 (6):640–7.

10 Wyatt MRR, Pathak SB, Zibarras L. Advancing selection in an SME: is best practice methodology applicable? Int Small Bus J 2010;28 (3):258–73.

11 Clevenger J, Pereira GM, Wiechmann D, Schmitt N, Harvey VS. Incremental validity of situational judgement tests. J Appl Psychol 2001;86 (3):410–7.

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

13 McDaniel M, Whetzel DL, Schmidt F, Maurer SD. The validity of employment interviews: a comprehensive reviewandmeta-analysis.JApplPsychol1994;79(4):599–616.

14 Steiner DD, Gilliland SW. Fairness reactions to personnel selection techniques in France and the United States. J Appl Psychol 1996;81 (2):134–41.

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

16 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;96 (5):927–40.

17 Plint S, Patterson F. Identifying critical success factors for designing selection processes into postgraduate specialty training: the case of UK general practice. Postgrad Med J 2010;86 (1016):323–7.

18 Lievens F, Buyse T, Sackett P. The operational validity of a video-based situational judgement test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. J Appl Psychol 2005;90 (3):442–52.

19 McDaniel M, Whetzel D. Situational judgement tests. In: Whetzel DL, Wheaton GR, eds. Applied Measurement: Industrial Psychology in Human Resources Management. Mahwah, NJ: Lawrence Erlbaum Associates 2007; 235–57.

20 McDaniel M, 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.

21 Rust J, Golombok S. Modern Psychometrics: The Science of Psychological Assessment. London: Routledge 1999.

22 Ployhart R, Ehrhart MG. Be careful what you ask for: effects of response instructions on the construct validity and reliability of situational judgement tests. Int J Select Assess 2003;11 (1):1–16.

23 McDaniel M, Nguyen NT. Situational judgement tests: a review of practice and constructs assessment. Int J Select Assess 2001;9:103–13.

24 Ployhart R, Weekley JA, Holtz BC, Kemp C. Web-based and paper-and-pencil testing of applicants in a proctored setting: are personality, biodata, and situational judgement comparable? Pers Psychol 2003;56: 733–52.

25 Chan D, Schmitt N. Situational judgement and job performance. Hum Perform 2002;15 (3):233–54.

26 Lievens F, Buyse T, Sackett P. Retest effects in operational selection settings: development and test of a framework. Pers Psychol 2005;58 (4):981–1007.

27 Bledow R, Frese M. A situational judgment test of personal initiative and its relationship to performance. Personnel Psychology. 2009;62 (2):229–258.

28 Borman WC, Hanson MA, Oppler SH, Pulakos ED, White LA. Role of early supervisory experience in supervisor performance. J Appl Psychol 1993;78 (3): 443–9.

29 Gaugler BB, Rosenthal DB, Thornton GC, Bentson C. Meta-analysis of assessment centre validity. J Appl Psychol 1987;72 (3):493–511.

30 Rothstein HR, Schmidt F, Erwin FW, Owens WA, Sparks CP. Biographical data in employment selection: can validities be made generalisable? J Appl Psychol 1990;75 (2):175–84.

31 McDaniel M, Hartman N, Whetzel D, Grubb W. Situational judgement tests, response instructions, and validity: a meta-analysis. Pers Psychol 2007;60 (1): 63–91.

32 O'Connell MS, Hartman NS, McDaniel MA, Grubb WL III, Lawrence A. Incremental validity of situational judgement tests for task and contextual job performance. Int J Select Assess 2007;15 (1):19–29.

33 Goldstein HW, Zedeck S, Goldstein IL. g: is this your final answer? Hum Perform 2002;15:123–42.

34 Jansen PGW, Stoop BAM. The dynamics of assessment centre validity: results of a 7-year study. J Appl Psychol 2001;86 (4):741–53.

35 Mullins ME, Schmitt N. Situational judgement testing: will the real constructs please present themselves? Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, 23–26 April 1998, Dallas, TX.

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

37 Smith KC, McDaniel M. Criterion and construct validity evidence for a situational judgement measure. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, 23–26 April 1998, Dallas, TX.

38 Koczwara A, Patterson F, Zibarras L, Kerrin M, Irish B, Wilkinson M. Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. Med Educ 2012;46:399–408.

39 Schmidt F. The future of personnel selection in the US Army. In: Weekley JA, Ployhart RE, eds. Situational Judgment Tests. Mahwah, NJ: Lawrence Erlbaum Associates 1994;333–50.

40 Tesluk PE, Jacobs RR. Toward an integrated model of work experience. Pers Psychol 1998;51 (2):321–55.

41 Quin˜ones MA, Ford JK, Teachout MS. The relationship between work experience and job performance: a conceptual and meta-analytic review. Pers Psychol 1995;48 (4):887–910.

42 Weekley JA, Jones C. Further studies of situational tests. Pers Psychol 1999;52 (3):679–700.

43 Patterson F, Denney M, Wakeford R, Good D. Making assessment fair and equal in postgraduate training: an interdisciplinary perspective of a future research agenda. Br J Gen Pract 2011;61 (593):712–3.

44 Motowidlo SJ, Tippins N. Further studies of the low-fidelity simulation in the form of a situational inventory. J Occup Organ Psychol 1993;66:337–44.

45 Oswald FL, Schmitt N, Kim BH, Ramsay LJ, Gillespie MA. Developing a biodata measure and situational judgement inventory as predictors of college student performance. J Appl Psychol 2004;89 (2):187–207.

46 Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK-trained doctors and medical students: systematic review and meta-analysis. BMJ 2011;342:901.

47 Potosky D, Bobko P, Roth PL. Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: corrected estimates and realistic expectations. Int J Select Assess 2005;13 (4):304–15.

48 Schmidt F. The problem of group differences in ability test scores in employment selection. J Vocat Behav 1988;33 (3):272–92.

49 Nguyen NT, Biderman MD, McDaniel M. Effects of response instructions on faking a situational judgement test. Int J Select Assess 2005;13 (4):250–60.

50 Chan D, Schmitt N. Video-based versus paper-and-pencil method of assessment in situational judgement tests: subgroup differences in test performance and face validity perceptions. J Appl Psychol 1997;82 (1):143–58.

51 Anderson N, Lievens F, Van Dam K, Marise B. A construct-driven investigation of gender differencesin a leadership-role assessment centre. J Appl Psychol 2006;91 (3):555–66.

52 Gilliland SW. The perceived fairness of selection systems: an organisational justice perspective. Acad Manag Rev 1993;18 (4):694–734.

53 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.

54 Hausknecht JP, Day DV, Thomas SC. Applicant reactions to selection procedures: an updated model and meta-analysis. Pers Psychol 2004;57 (3):639–83.

55 Lievens F, Sackett P. Situational judgement tests in high-stakes settings: issues and strategies with generating alternate forms. J Appl Psychol 2007;92 (4):1043–55.

56 Kanning UP, Kuhne S. Social desirability in a multimodal personnel selection test battery. Eur J Work Organ Psychol 2006;15 (3):241–61.

57 Patterson F, Zibarras L, Carr V, Irish B, Gregory S. Evaluating candidate reactions to selection practices using organisational justice theory. Med Educ 2011;45 (3):289–97.

58 Cullen M, Sackett P, Lievens F. Threats to the operational use of situational judgement tests in the college admission process. Int J Select Assess 2006;14 (2):142–55.

59 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.

60 Jensen AR. The g Factor: Science of Mental Ability (Human Evolution, Behavior & Intelligence). Westport, CT: Greenwood Press 1998.

61 Costa PT, McCrae RR. NEO PI-R professional manual. Odessa, FL: Psychological Assessment Resources 1999.