2015, 1–15, Early Online MEDICAL TEACHER

AMEE GUIDE

Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100

FIONA PATTERSON[1,2], LARA ZIBARRAS[3] & VICKI ASHWORTH[1]

[1]Work Psychology Group, UK, [2]University of Cambridge, UK, [3]City University London, UK

Abstract

Why use SJTs? Traditionally, selection into medical education professions has focused primarily upon academic ability alone. This approach has been questioned more recently, as although academic attainment predicts performance early in training, research shows it has less predictive power for demonstrating competence in postgraduate clinical practice. Such evidence, coupled with an increasing focus on individuals working in healthcare roles displaying the core values of compassionate care, benevolence and respect, illustrates that individuals should be selected on attributes other than academic ability alone. Moreover, there are mounting calls to widen access to medicine, to ensure that selection methods do not unfairly disadvantage individuals from specific groups (e.g. regarding ethnicity or socio-economic status), so that the future workforce adequately represents society as a whole. These drivers necessitate a method of assessment that allows individuals to be selected on important non-academic attributes that are desirable in healthcare professionals, in a fair, reliable and valid way.

What are SJTs? Situational judgement tests (SJTs) are tests used to assess individuals' reactions to a number of hypothetical role-relevant scenarios, which reflect situations candidates are likely to encounter in the target role. These scenarios are based on a detailed analysis of the role and should be developed in collaboration with subject matter experts, in order to accurately assess the key attributes that are associated with competent performance. From a theoretical perspective, SJTs are believed to measure prosocial Implicit Trait Policies (ITPs), which are shaped by socialisation processes that teach the utility of expressing certain traits in different settings such as agreeable expressions (e.g. helping others in need), or disagreeable actions (e.g. advancing ones own interest at others, expense).

Are SJTs reliable, valid and fair? Several studies, including good quality meta-analytic and longitudinal research, consistently show that SJTs used in many different occupational groups are reliable and valid. Although there is over 40 years of research evidence available on SJTs, it is only within the past 10 years that SJTs have been used for recruitment into medicine. Specifically, evidence consistently shows that SJTs used in medical selection have good reliability, and predict performance across a range of medical professions, including performance in general practice, in early years (foundation training as a junior doctor) and for medical school admissions. In addition, SJTs have been found to have significant added value (incremental validity) over and above other selection methods such as knowledge tests, measures of cognitive ability, personality tests and application forms. Regarding differential attainment, generally SJTs have been found to have lower adverse impact compared to other selection methods, such as cognitive ability tests. SJTs have the benefit of being appropriate both for use in selection where candidates are novices (i.e. have no prior role experience or knowledge such as in medical school admissions) as well as settings where candidates have substantial job knowledge and specific experience (as in postgraduate recruitment for more senior roles). An SJT specification (e.g. scenario content, response instructions and format) may differ depending on the level of job knowledge required. Research consistently shows that SJTs are usually found to be positively received by candidates compared to other selection tests such as cognitive ability and personality tests. Practically, SJTs are difficult to design effectively, and significant expertise is required to build a reliable and valid SJT. Once designed however, SJTs are cost efficient to administer to large numbers of candidates compared to other tests of non-academic attributes (e.g. personal statements, structured interviews), as they are standardised and can be computer-delivered and machine-marked.

Introduction

Traditionally, students and trainees within medicine have tended to be assessed on academic ability alone. This approach however has a number of limitations, and recently there has been an increasing emphasis in medical education and training on assessing for non-academic and professional attributes that are important for competent performance in clinical practice (Eva et al. 2009; Prideaux et al. 2011). In the UK for example, there are practical limitations of selecting based on the academic ability alone, since selecting students based on their A levels is now problematic due to grade inflation (James et al. 2010) with reduced discriminatory power

Correspondence: Professor Fiona Patterson, Department of Psychology, University of Cambridge and Work Psychology Group, 27 Brunel Parkway, Pride Park, Derby DE24 8HR, UK. Tel: +44 1332 295687; E-mail: f.patterson@workpsychologygroup.com

ISSN 0142-159X print/ISSN 1466-187X online/15/0000001–15 © 2015 Taylor & Francis
DOI: 10.3109/0142159X.2015.1072619

Practice points

- Situational judgement tests (SJTs) provide a reliable method for measuring important non-academic attributes (such as empathy, integrity and teamwork) that are important for education, training and practice in medicine and a wide range of healthcare roles.
- SJTs are a measurement methodology rather than a single test *per se*. As such, SJTs can differ markedly from each other (in scenario content and response formats, for example), and should be designed in collaboration with subject matter experts (SMEs) to ensure their relevance, appropriateness and fairness regarding the target role.
- SJTs measure prosocial Implicit Trait Policies (ITPs), which are beliefs about the cost or benefits of acts expressing compassion, caring and respect for patients, related to candidates' trait expression and values.
- When designed appropriately, compared to other selection tools, SJTs are generally reliable, valid, fair and well received by candidates.
- Although good quality SJTs are difficult to design, once developed, SJTs are a cost-effective and an efficient selection tool for non-academic attributes as they can be computer-delivered and machine-marked.

due to a large proportion of medical candidates attaining top grades (McManus et al. 2008).

Although academic attainment has been shown to be a good predictor of early performance in medical school (Ferguson et al. 2002), longitudinal research shows that the predictive power of academic attainment declines as trainees move into clinical practice and beyond (Ferguson et al. 2002; James et al. 2010; McManus et al. 2013). These findings emphasise that academic ability is a necessary, but not sufficient condition to ensure that trainees become competent health professionals, and thus the importance of selecting trainees on a range of non-academic attributes in addition.

Within the UK, although the values and behaviours expected of health and social care professionals are preserved in the National Health Service (NHS) Constitution (2013), recent government enquiries (Cavendish 2013; Francis 2013) have highlighted major concerns about the decline in compassionate care within all healthcare roles, which has global significance. These enquiries, although UK-based, have relevance internationally, as they highlight the critical role that the workforce plays in ensuring the provision of high quality and safe healthcare services and, in particular, the impact of staff values and behaviours on the quality of patient care and thus patient outcomes. Undoubtedly, an important first step is ensuring that the right individuals with the appropriate values to work in clinical practice are appointed to any educational course, training place or healthcare role.

Job analysis research provides supporting evidence for the importance of non-academic attributes for successful performance in various healthcare roles. For example, attributes such as empathy, resilience, team involvement and integrity are necessary for medicine and dental students (Patterson et al. 2012a) and in postgraduate medical training (Patterson et al. 2000, 2008). In terms of selection, this presents a challenge regarding how to reliably assess values, personal qualities and attributes in an effective and efficient way, because many methods that aim to assess non-academic attributes have not been found to be robust (Albanese et al. 2003). For example, there is a substantial body of evidence that suggests that traditional methods of screening candidates on the basis of non-academic attributes, such as personal statements and references, do not provide valid assessments of candidates' ability or non-academic attributes (Ferguson et al. 2000, 2003; Poole et al. 2009; Kreiter & Axelson 2013; Husbands et al. 2014; Patterson et al. 2016). Regarding personality assessments, although evidence suggests that they can provide added value in selection, (especially when used to complement structured interviews), some researchers have expressed caution in using personality tests as screening tools for high stakes selection (such as medicine), due to the risk of faking and susceptibility to coaching (Rankin 2013). In contrast, SJTs can be designed to be less susceptible to coaching (Patterson et al. 2013a) and recent research suggests that access to coaching does not materially influence the operational validity of SJTs in practice (Stemig et al. 2015).

A systematic review of the research evidence shows that SJTs are a valid and reliable method for assessing non-academic attributes (Patterson et al. 2012b) in the context of medical education and training, and the evidence is explored in detail in this Guide.

Aim/objective of the guide

Our aim is to provide an overview of the evidence-base for SJTs, describing how SJTs compare to other assessment tools, and how and where SJTs can be used. Specifically, the evidence for their use in selection will be explored, focusing on reliability, validity, fairness, susceptibility to coaching and the theory underpinning SJTs. We explore the practical aspects of developing SJTs, including item structure, format, response options and scoring. We also provide several illustrative examples and case studies alongside some comments about future SJT research and implications for practice in medical education and training. Although this Guide frequently references UK practice, other international organisations and institutions can nonetheless use the content and principles.

What are situational judgement tests?

It is important to note that SJTs are a *measurement methodology* rather than a single style of assessment, as the scenario content, response instructions, response formats and approaches to scoring can vary significantly across settings. Typically, candidates sitting an SJT are presented with hypothetical written or video-based scenarios of a situation they are likely to encounter in the given role. Candidates are asked to identify the appropriateness or effectiveness of various

Box 1. An example SJT item for postgraduate medical education.

On the morning ward round, your registrar/specialty trainee said that Mrs Anderson is medically fit following her total knee replacement and could be discharged if Occupational Therapy feel it is appropriate. The occupational therapist has assessed Mrs Anderson and believes it is safe for her to go home with a care package that has been arranged. It is now 4 p.m. and the nurse informs you that Mrs Anderson is demanding to see a doctor, as she does not feel that she is ready to go home yet. An elective admission is waiting in the day room for Mrs Anderson's bed.
Rank in order the appropriateness of the following actions in response to this situation (1 = Most appropriate; 5 = Least appropriate).
A. Ask Mrs Anderson about her concerns.
B. Ask a senior colleague to speak with Mrs Anderson.
C. Ask the bed manager if he can find another bed for the elective patient.
D. Explain to Mrs Anderson that the bed has already been allocated and she has to go home.
E. Ask the occupational therapist to come and speak to Mrs Anderson with you.

Box 2. An example SJT item for selection into specialty training.

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 appropriateness of the following actions in response to this situation.
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.

response options from a pre-defined list of alternatives (Boxes 1–3). These response options are designed in advance with a pre-determined scoring key agreed by subject matter experts. A single SJT is likely to comprise several scenarios so that a broad range of constructs can be measured efficiently.

An SJT's content, format and test length is designed to fit the role, the selection criteria and specification requirements for the test (Lievens et al. 2008). SJTs have been used in assessment for over 40 years across a broad range of occupational contexts in both the public and private sectors (Chan et al. 1998; Ployhart et al. 2003; Wyatt et al. 2010), and more recently have been applied to roles in medicine and other healthcare professions – see Patterson et al. (2016) for a systematic review.

Role analysis

Whether designed for selection, assessment or development purposes, to follow best practice, SJT scenarios should be based on a thorough analysis of the relevant role in order to assess the key attributes and competencies that are associated with competent performance in the role. This ensures that the content of the SJT reflects work-, education- or training-related situations that candidates are likely to encounter in the target role. In addition, the test specification should be developed in collaboration with key stakeholders and current role incumbents (Motowidlo et al. 1990), and it is important for a thorough design and evaluation process to be undertaken to ensure the psychometric quality of an SJT (Patterson et al. 2009b). A detailed guide to the steps of developing an SJT in line with best practice is provided towards the end of this Guide.

Response instructions and format

Response instructions for SJTs typically fall into one of two categories: knowledge based (i.e. *what is the best option*) or behavioural tendency (i.e. *what would you be most likely to do*). A variety of response formats can be used within each of these categories, such as *ranking* possible actions in order, *rating* all response options independently, choosing the three best answers from a larger number of response options (*multiple choice*), or choosing the best and/or worst response options. Some researchers have developed a single-response SJT format, whereby only one response option is given as part of the scenario (Motowidlo et al. 2009; Martin & Motowidlo 2010).

The type of response format used depends on the test specification and the context or level in the education and training pathway that the SJT is targeting, as different response formats are differentially related to knowledge, cognitive ability and other constructs. For example, for medical school admissions, this is essentially a novice population and we cannot (and should not) assume job specific knowledge or experience. In contrast, SJTs for postgraduate selection do assume some job experience to retain scenario authenticity and validity. Response alternatives can be presented in either a written (low fidelity) or a video-based (medium fidelity) format (Lievens et al. 2008; Christian et al. 2010) and each approach has both advantages and disadvantages (relating to aspects such as candidate reactions, validity and cost, for example). The reliability and effectiveness of different item presentation formats and response instructions of SJTs will be explored in this Guide.

SJTs are scored by comparing candidates' responses to a pre-determined scoring key, which dictates the scores obtained for each answer, and has previously been agreed

via an in-depth review process by a group of subject matter experts (SMEs). In Box 2, we outline an example SJT item for the role of a junior doctor for selection into specialty training. Here, the response options are knowledge oriented (i.e. what should you do/what is the best option) and the response format asks candidates to rank all possible responses in order of appropriateness, which represents a relatively complex set of judgements.

In this scenario, you will note that the test taker is already told there has been a prescribing error so the aim is not to test clinical knowledge *per se*. Instead, the focus of the item relates to a set of interpersonal dilemmas. In this scenario, the most appropriate (first thing to do) is answer B – to correct the prescription. However, there are then a complex set of judgements for a candidate to make regarding how to best deal with some challenging interpersonal dilemmas, i.e. how best to deal with the consultant, the nurses, co-workers and indeed the patient. In this way, by presenting several scenarios in a test, SJTs can measure a broad range of professional attributes.

Theory behind SJTs – How do they work?

Historically, researchers have engaged in considerable debate regarding the construct validity of SJTs (i.e. what do SJTs measure?), but there is now a relatively clear picture of the theoretical underpinnings of SJTs.

SJTs are based on two key theoretical propositions. First, SJTs are derived from a long established *behavioural consistency* theory (i.e. that past behaviour is the best predictor of future behaviour), in which the central principle is that eliciting a sample of current behaviour allows the prediction of future (i.e. in-role) behaviour (Wernimont & Campbell 1968; Motowidlo et al. 2006).

Second, there is a growing consensus in the research literature that SJTs measure prosocial *Implicit Trait Policies* (ITPs), and depending on the job level, specific job knowledge, as in postgraduate selection (Motowidlo & Beier 2010; Patterson et al. 2015b). ITP theory proposes that individuals develop beliefs about the effectiveness of different behaviours – that is, beliefs about the costs and benefits associated with expressing certain traits (which guide behaviours) in particular situations, in relation to individuals' inherent tendencies or traits. As individuals make judgements about how and when to express certain traits, ITPs are related to choices about trait *expressions* rather than traits *per se* (Motowidlo et al. 2006). For example, a doctor dealing with a sensitive situation in the workplace (such as the death of a relative) may have to make a judgement about the utility (costs/benefits) of demonstrating empathy and agreeableness as a more successful strategy than acting brusquely or being disagreeable (even if the doctor's preference tends towards being generally disagreeable and/or empathetic). As another example, imagine an anaesthetist whose personality profile shows a preference for introversion, but decides to act in what might be observed as an extraverted manner, as this is the most effective approach for the situation at hand.

ITPs are thought to be shaped by experiences during socialisation processes, such as through parental modelling throughout childhood (or later during tutoring and role models at medical school). This may teach individuals the utility of, for example; agreeable expressions, that is, helping others in need, or turning the other cheek; or disagreeable expressions, that is, showing selfish pre-occupation with one's own interests, holding a grudge/'getting even' or advancing ones own interests at another person's expense. As such, SJTs may also represent a promising tool for assessing individuals' values, as the element of personal choice involved when behaving consistently with one's values can be measured by an SJT (i.e. candidates must make a choice about which response options are the most appropriate or important) (Parks & Guay 2009; Patterson et al. 2015b).

Research evidence suggests that SJTs are effective predictors of job performance because SJTs measure procedural awareness about effective behaviour in a given situation (including domain-based knowledge where appropriate), and, relatedly, individuals' beliefs about the costs/benefits of expressing personality traits in role-related situations (Motowidlo et al. 2006; Lievens & Patterson 2011).

What is the research evidence for SJTs?

Any selection method in operational use must meet exacting standards relating to the psychometric properties (i.e. reliability, validity, accuracy), in addition to being acceptable to stakeholders, in that particular context (Schmidt & Hunter 1998; Arnold et al. 2010; Patterson et al. 2012c). These standards ensure a method provides an effective, acceptable and legally defensible means of selection, which is arguably especially important within high-stakes, high-volume selection contexts such as in medicine. As such, when SJTs are used in medicine (and in other settings), they should be regularly reviewed and evaluated to ensure that they meet these psychometric and acceptability requirements. The research evidence for SJTs relating to these issues is reviewed in the following sections.

How and when can SJTs be used?

SJTs may be used for selection, assessment or development, and have the benefit of being designed in a way that is tailored to fit the specific purpose and needs of the target role. In medicine, SJTs are often used as a screening tool during selection as they enable the assessment of the non-academic abilities and attributes of large numbers of candidates in a standardised and cost-efficient way (Koczwara & Ashworth 2013). Those who successfully complete an SJT are then typically shortlisted to the next stage of the selection process, such as a structured interview.

SJTs can also be used effectively at the interview stage, as one of several stations of a multiple mini-interview, for example. As such, SJTs tend to form one part of a multi-method selection process, whereby candidates are assessed across a range of attributes and skills, as identified by the role analysis. A large number of scenarios can be

measured in an SJT, whereas in interviews, a small number of scenarios can be discussed and probed, such that SJTs and structured interviews are complementary.

In the UK, SJTs are increasingly being incorporated into healthcare selection, with many recent examples in medical education and training. For example, an SJT is being used alongside a battery of cognitive ability tests for UK medical schools admissions, demonstrating promising evidence of reliability and validity (Patterson & Martin 2014; Patterson et al. 2014; Sartania et al. 2014).

Alongside other methods, an SJT measuring empathy, integrity and resilience has been used successfully for selection into training into UK general practice for over 10 years, which has also demonstrated good predictive validity, positive candidate reactions and significant cost savings regarding the resources required for hand-marking personal statements, for example (Patterson et al. 2009a; Plint & Patterson 2010; Lievens & Patterson 2011).

An SJT has been used alongside an educational performance measure to select all doctors applying for UK Foundation training since 2013, also demonstrating predictive validity with subsequent training outcomes and job performance. This SJT is designed to assess for the key attributes identified in a role analysis (including commitment to professionalism, coping with pressure, effective communication, patient focus, and working effectively as part of a team), and has performed well psychometrically year on year.

In the UK, SJTs are used for postgraduate training selection for a variety of roles including Public Health (Kerrin et al. 2014), Psychiatry (Lopes et al. 2015a) and Ophthalmology (Lopes et al. 2015b). SJTs are also used for a variety of other healthcare professions including dental foundation training (Patterson et al. 2012a) and veterinary science (Kerrin et al. 2015).

The use of SJTs in healthcare selection is expanding globally. Internationally, SJTs have been used in medical school admissions in Belgium (Lievens et al. 2005a), Singapore (Ashworth et al. 2014), Canada (Dore et al. 2009) and in postgraduate recruitment in Australia to select trainees for entry to training in general practice (Roberts & Togno 2011; Roberts et al. 2014). Many other countries are piloting the use of SJTs in admissions; for example, an SJT is being piloted at the American Association of Medical Colleges (AAMC – https://www.aamc.org/initiatives/admissionsinitiative/sjt/), alongside the MCAT.

How reliable are SJTs?

There are a number of ways to ensure that an SJT is *reliable* in measuring the constructs it aims to assess (based on the role analysis) in a consistent way. This includes measuring *internal consistency,* which assesses whether several items that should measure the same construct produce similar scores; *test–retest reliability* which indicates whether a measure is stable over a given time period; and *parallel forms reliability* where two versions of the same test correlate sufficiently (Rust & Golombok 1999).

SJTs have been described as *construct heterogeneous* at the item level (McDaniel & Whetzel 2007) since any one item may target several performance dimensions or constructs (in other words, SJTs rarely measure one single dimension), which implies an inherent difficulty in accurately estimating the internal consistency of an SJT, and a risk of underestimating their reliability (McDaniel & Whetzel 2007; Catano et al. 2012). Indeed, a meta-analysis of 39 different studies found Cronbach's alpha coefficients ranging from $\alpha = 0.43$–0.94 (McDaniel et al. 2001). However, internal consistency of SJTs used in medical and dental contexts (Koczwara et al. 2012; Patterson et al. 2012a, 2014, 2015a) have been found to consistently approach or exceed $\alpha = 0.70$; the accepted value indicating good internal consistency (Kline 2000).

Some researchers argue that *test–retest* or *parallel forms reliability* may be a more accurate approach to examining the reliability of an SJT since internal consistency is more appropriate for uni-dimensional (*construct homogenous*) tests (McDaniel & Nguyen 2001; Lievens et al. 2008; Catano et al. 2012). Research shows a range of values for test-retest reliability of SJTs, from $r = 0.20$ to $r = 0.92$ (Ployhart & Ehrhart 2003; Ployhart et al. 2003), dependent upon the quality of the test construction.

Parallel forms reliability is measured by the correlation of two tests that are assessing the same construct (Rust & Golombok 1999). Kline (2000) recommends that parallel forms reliability should be above $r = 0.70$. On balance, the research evidence shows good levels of parallel forms reliability for SJTs, for example: $r = 0.76$ (Chan & Schmitt 2002) and $r = 0.66$ (Lievens et al. 2005b).

In summary, the research broadly shows that SJTs have moderate to good levels of reliability, regardless of the type of method used to assess reliability. Since SJTs are a measurement method and there is no one single type of SJT, it is important to evaluate each test independently to judge reliability. When evaluating reliability, internal consistency may sometimes be a less appropriate measurement of SJT consistency than other measures of reliability, due to the usually heterogeneous nature of SJT items.

How valid are SJTs?

As with all other assessment tools, there are multiple approaches to evaluate the *validity* of an SJT (i.e. is the tool measuring the criteria that it purports to measure?). To establish the validity of a tool this includes measuring *criterion-related validity* (i.e. evaluating the extent to which scores on an SJT can predict subsequent performance or work behaviour); *incremental validity* (i.e. does the tool add value and predict variance in outcome criteria beyond other tools); and *construct validity* (i.e. evaluating the extent to which a tool is measuring what it is supposed to measure, where comparisons are made with measures of similar constructs).

Criterion-related validity

A central consideration when assessing the criterion-related validity of any selection method is to consider precisely *what the method is intended to measure*. For example, Lievens et al. (2005a) acknowledged the need to attend to the constructs underlying both predictors and criterion (outcome) measures when assessing relationships between the two. Considering this

in the context of SJTs, which are designed to assess non-academic constructs and interpersonal skills, we would not theoretically expect an SJT to predict performance on a highly cognitively loaded criterion. We would, however, expect SJTs to predict performance on criterion-matched outcomes, such as interpersonal skills courses, educational supervisor ratings of empathy and integrity or communication stations in an Objective Structured Clinical Examination (OSCE).

The literature regarding the criterion-related validity of SJTs in healthcare and other professions supports this proposition. For example, Lievens et al. (2005a) examined the criterion-related validity of an SJT for medical college admissions in Belgium, and found that the SJT showed incremental validity over cognitively oriented measures for curricula that included interpersonal courses, but not for other, academic curricula. Similarly, Oswald et al. (2004) found that an SJT predicted performance on dimensions such as leadership and perseverance, but did not predict grade point average, and Libbrecht et al. (2014) found that an SJT measuring emotional intelligence predicted performance in courses on communication and interpersonal sensitivity in medical school, beyond cognitive ability and conscientiousness.

Meta-analyses have found favourable criterion-related validity of SJTs compared to other selection methods. For example, Christian et al. (2010) found corrected operational validity coefficients of r=0.25–0.47, and McDaniel et al. (2001) estimated a criterion-related validity of $r = 0.34$ across multiple occupational groups. In addition, SJTs have been shown to have comparable or increased predictive validity over structured interviews (McDaniel et al. 1994), selection centres (Gaugler et al. 1987) and cognitive ability tests (Clevenger et al. 2001).

Regarding medical education specifically, on-going evaluation has shown that an SJT used in selection for postgraduate training in the UK predicts subsequent in-training performance and end-of-training competence after three years, with large effect sizes (Patterson et al. 2013b). Longer-term studies also show that the SJT predicts performance in training in relation to educational supervisors' ratings (Lievens & Patterson 2011) and performance in clinical skills licensing OSCEs (Patterson et al. 2012c).

Two longitudinal, multiple cohort studies in Belgium found an SJT used within medical school admissions predicted a number of assessments at different points throughout training, in addition to job performance nine years later in both studies: $r = 0.22$ (Lievens & Sackett 2012) and $r = 0.15$ (Lievens 2013).

In summary, SJTs have been found to have good levels of predictive validity across a wide range of different occupations, including in healthcare education and training. Methodologically, it is important to ensure that appropriate outcome criteria and measures are identified in order to meaningfully assess the validity of SJTs.

Incremental validity

It is important to establish whether an SJT has predictive validity over and above other selection methods, such as cognitive ability, personality and job knowledge tests, to assess the value added by an SJT in a selection system (i.e. what is the added benefit of using an SJT?). The presence or absence of such added (incremental) validity can provide evidence for the cost-effectiveness and efficiency of SJTs, as well as to support their theoretical rationale.

The incremental validity of SJTs has been widely demonstrated in the literature across numerous operational groups, including healthcare. SJTs explained an additional 2.6% of variance over the combined predictive validity of conscientiousness, job experience and job knowledge in federal investigator officers' job performance (Clevenger et al. 2001). In numerous manufacturing organisations, an SJT showed incremental validity over cognitive (3%) and personality (4%) measures for the prediction of task performance, and over cognitive ability (4%) when predicting contextual performance (O'Connell et al. 2007).

In the context of medicine specifically, Lievens et al. (2005a) found that an SJT assessing interpersonal awareness for entry into medical schools in Belgium had incremental validity over the cognitive measures for components of the curricula that had an interpersonal skills component. Patterson et al. (2009a) showed that an SJT used for selection in postgraduate general practice training had substantial incremental validity over scores from an application form (17%) and a job-knowledge test (14%). Similar findings have been demonstrated for an SJT into Core Medical Training in the UK (Patterson et al. 2009b).

In summary, there is a wealth of evidence in healthcare and beyond that SJTs provide added value over other selection methods for predicting job performance.

Considerations in assessing the predictive validity of SJTs

Research suggests that the predictive validity of measures of non-academic attributes differs over time in the context of medical education and training (i.e. predictive validity is higher when students enter clinical practice). We propose that this is because during the classroom-based, didactic setting in which learning takes place during the early years of a medical degree, the focus is more on cognitive indicators and academic attainment. However, as students move into clinical practice and beyond, the necessity for interpersonal skills comes to the fore because the team-based working and patient contact are key measures of competence (Ferguson et al. 2014). This proposition is supported by other researchers (McDaniel et al. 2007; Lievens et al. 2008), and reiterates the importance of choosing not only appropriate outcome criteria to assess the validity of SJTs, but also of collecting outcome data at the appropriate time.

Emerging evidence from the medical education literature in the UK indicates that SJTs may have greater predictive validity at the lower end of performers (as yet unpublished research by the present authors across a number of SJTs for medical admissions in the UK). For example, applicants lacking prosocial ITPs are less likely to perform well on an SJT. Further research both internally and in broader healthcare contexts is required to replicate and extend these findings.

In summary, it is important to take into consideration the point in a healthcare education or training pathway that outcome data is collected, in order to most accurately assess

the validity of an SJT. Emerging evidence suggests SJTs may be best used to 'select out' candidates as they are potentially better predictors at the lower end of the distribution of scores, however evidence for this is in its infancy.

Construct validity

As outlined previously, SJTs represent a methodology rather than a specific test. As such, scenarios are context-specific and responses can be designed to assess a range of different attributes or constructs. Consequently, SJTs are designed to measure a wide range of different attributes, and a candidate's response may represent a combination of their ability, personality and experience (Lievens et al. 2008). Therefore, due to the heterogeneous nature of different SJTs, with a broad range of different test specifications available, it is challenging for researchers to describe specifically what SJTs as a methodology measure (i.e. their construct validity), and a range of constructs have been identified as being associated with scores on SJTs.

McDaniel et al.'s (2001) review shows that SJTs correlate with cognitive ability, although the magnitude of the relationship depended on several factors. SJTs based on a job analysis correlated more highly with cognitive ability ($r = 0.50$) than those that were not ($r = 0.38$), and SJTs which contained more detailed questions were more highly correlated with cognitive ability ($r = 0.56$) than those less detailed ($r = 0.47$). It is suggested that SJTs that are cognitively oriented (such as those involving planning, organising and problem-solving) tend to correlate higher with cognitive ability than those oriented solely on interpersonal issues such as empathy. In postgraduate medical selection, Koczwara et al. (2012) found that an SJT correlated significantly with cognitive ability tests, suggesting that they measure overlapping constructs. In contrast, Clevenger et al. (2001) found SJT scores were not correlated with cognitive ability, and an SJT used for assessment of professionalism in dental school based on a role analysis did not correlate with a test of clinical knowledge (Escudier et al. 2015). As such, the evidence indicates that the extent to which SJTs measures cognitive ability varies, depending on the specific design of each individual SJT.

In relation to personality, Mullins and Schmitt (1998) found that an SJT was most strongly correlated with the personality traits of Conscientiousness ($r = 0.26$) and Agreeableness ($r = 0.22$). A meta-analysis found SJT scores had a corrected correlation of $r = 0.25$ with Agreeableness; $r = 0.31$ with Emotional Stability, $r = 0.26$ with Conscientiousness, $r = 0.06$ with Extraversion and $r = 0.09$ with Openness (McDaniel & Nguyen 2001). McDaniel et al. (2007) found similar results, indicating that SJTs correlate most highly with the personality traits of Agreeableness and Conscientiousness, i.e. cooperation, forgiveness and a tendency to defer to others in interpersonal conflict; and achievement striving, leadership skills and organisation, respectively (McCrae & Costa 2008).

Motowidlo (2003) suggests the relationship between personality and ITPs (described earlier) can be explained via the notion of *dispositional fit*; that is, some individuals' dispositions (i.e. their personalities) are more appropriate for handling challenging social situations than others' dispositions. Dispositional fit specifies that people develop beliefs consistent with their personality traits about how best to handle difficult situations. Situations encountered in any given role are believed to vary in the extent to which they require behavioural responses that are expressive of a given trait. When individuals' beliefs about the appropriate behavioural responses to a situation correspond to the type of response actually required by the situation, those individuals possesses greater knowledge because their beliefs are correct (Campbell et al. 1996). As people often believe that actions expressive of their own personalities are most effective, people whose traits match those demanded by the situation will be most likely to possess knowledge of how to effectively behave in that situation.

Considering the relationship between SJT scores and Agreeableness, for example, because agreeable people should have ITPs that weigh Agreeableness more strongly, they are more likely to discriminate accurately between SJT response options, according to the level of Agreeableness expressed by each response option; thereby achieving higher scores on an SJT.

The literature has not yet fully converged on the association between SJTs and role knowledge. Earlier research has suggested that SJTs may simply be tests of role-specific knowledge (Schmidt 1994), whilst others suggest that since *experience* is multifaceted, different operationalisations of experience will result in different relationships between experience in a role and performance on an SJT. Indeed, Motowidlo and Beier (2010) posit that the type of experience that leads to the development of ITPs (i.e. general experience), is different to the type of experience that allows individuals to develop specific knowledge about a particular role. They conclude that the ITPs and specific role knowledge combine to produce the procedural knowledge measured by an SJT, but they have independent effects on role performance. Moreover, some SJTs are more strongly related to general experience (ITPs) than to highly role-specific experience.

In summary, since SJTs can be designed to assess a range of different attributes, research regarding their construct validity has been mixed. Research identifies that SJTs can correlate with cognitive ability, personality and knowledge, depending on the specification of the SJT.

What is the reliability and validity of different SJT formats, response formats and instructions?

Response formats

Research has focused on the relative benefits of response formats based on knowledge based (i.e. *what is the best option*) or behavioural tendency (i.e. *what would you be most likely to do*) formats (McDaniel & Nguyen 2001; McDaniel et al. 2003, 2007). St-Sauveur et al. (2014) posit that an argument in favour of a knowledge based format is that regardless of whether an SJT uses a knowledge based or behavioural tendency format, it will still measure the extent to which candidates know what the "correct" behaviour is in a given situation (i.e. knowing *what you should do*). In addition, they argue that knowledge based response formats lead to greater

certainty as to what the test measures (i.e. clearer construct validity). A meta-analysis (McDaniel et al. 2007) also found a better correlation with work performance for knowledge based response formats.

A key design consideration regarding response formats is the extent to which an SJT is required to be cognitively loaded, which will depend on its intended use and target candidate population. For example, it would be more appropriate for an SJT used for selection into a highly specialised trainee post (where specialist and clinical detail may be required to contextualise the scenarios) to have a greater degree of cognitive loading than an SJT for entry into an undergraduate healthcare degree. Whetzel and McDaniel (2009) conclude that SJTs with knowledge based instructions correlate more highly with cognitive ability than SJTs with behavioural tendency instructions. As such, test developers who wish to emphasise the assessment of personality constructs in an SJT may wish to use behavioural tendency instructions. On the other hand, if a selector wishes to maximise the variability in SJT scores based on cognitive ability within an SJT, a test with knowledge based instructions may be more appropriate (Whetzel & McDaniel 2009). Arguably, since medicine is a cognitively demanding profession, more cognitively loaded SJTs are more relevant.

Regarding construct validity and response instructions, Weekley et al. (2006) and McDaniel et al. (2007) found that SJT *format* may also influence construct validity: SJTs with knowledge-based instructions have higher correlations with cognitive ability than those with behavioural tendency instructions. Conversely, the researchers found that SJTs with behavioural tendency instructions correlated more strongly with personality traits. As such, SJTs with knowledge instructions may be considered measures of maximal performance, whilst behavioural tendency instructions may be measures of typical performance (Lievens et al. 2008).

Knowledge based SJTs however, if more cognitively loaded than behavioural tendency SJTs, are more likely to result in greater subgroup differences (Whetzel et al. 2008; Roth et al. 2013; St-Sauveur et al. 2014). Given that the reduction in subgroup differences is considered to be one of the key benefits of SJTs over other selection methods (compared to academic attainment), this may pose a significant challenge to test developers. Conversely, research shows behavioural tendency instructions are more susceptible to 'faking', similar to personality tests (Nguyen et al. 2005; Birkeland et al. 2006); and thus the balance of the costs and benefits of each response format must be considered when defining the specification of an SJT. Given that medical selection is competitive, it could be argued that knowledge based formats may be more appropriate for this high-stakes selection context, since SJTs are a measure of maximal performance (i.e. how a candidate performs at their peak), whereas behavioural tendency formats measure typical performance (i.e. how one typically behaves) (McDaniel et al. 2007). When knowledge based formats are used, evidence shows candidates are less able to "fake" the answer by attempting to give an answer that the candidate thinks the recruiter wants to hear. Similarly, research shows also that knowledge based formats are less susceptible to coaching than behavioural tendency formats (McDaniel et al. 2007), which is an important consideration in high stakes selection, such as medicine.

Response instructions

Research shows that the type of response *instructions* can influence the reliability of an SJT. For example, Ployhart and Ehrhart (2003) found that *rating the effectiveness of each response* results in the highest internal consistency ($\alpha = 0.73$), whilst *choosing the most effective response* results in the lowest internal consistency ($\alpha = 0.24$). St-Sauveur et al. (2014) found that a single best answer response format had the lowest internal consistency of SJT response formats, compared to rank-order and choosing the 'best and worst' responses. Similarly, Ployhart and Ehrhart (2003) found that test–retest stability depends on the type of response instructions used, with higher reliabilities when candidates are asked to rate the likelihood that they would perform each response option.

Outside of medicine, emerging evidence indicates promising validity of the single-response response SJT format as a measure of procedural knowledge and as a predictor of job performance (Motowidlo et al. 2009; Martin & Motowidlo 2010; Crook et al. 2011), however further research is required to ascertain the reliability and long-term validity of this response format.

In summary, there are advantages and disadvantages of all possible SJT response formats and instructions, which may significantly impact the reliability, and validity of an SJT; these are important considerations when defining the specifications and purpose of an SJT.

Box 3 shows three examples of SJT items using different types of response instructions. The choice of response instructions during test design reflects a number of considerations: the scenario content, the ability to provide and elicit the information needed, the target population (i.e. how cognitively loaded a response instruction should be), and the level of discrimination required between candidates. For example, the nature of some scenarios and the possible responses to them lend themselves to ranking items (requiring the ability to differentiate between singular responses to a scenario), whereas other scenarios lend themselves to multiple choice items (where it is necessary to do more than one thing, or tackle more than one aspect, in response to a scenario).

Format

Mostly, SJTs are delivered either as a written/text-based paper and pencil format or delivered online (Weekley & Ployhart 2005). It is also possible to deliver an SJT in a video-based format, which has the benefit of providing candidates with a more realistic, medium fidelity example of a role-related scenario (Sharma & Nagle 2015). Text based SJTs are significantly more cost effective to develop and maintain compared to video-based SJTs, and some evidence suggests that written SJTs have higher correlations with cognitive ability, due to the reading skills required (Chan & Schmitt 1997; Lievens & Sackett 2006). It could be argued that text based SJTs may be more appropriate in job roles that require advanced cognitive processing skills, such as medicine and other complex healthcare roles. Recent research has also

Box 3. Example items of situational judgement tests showing different response formats (Patterson et al. 2012b).

Multiple choice format

You review a patient on the surgical ward who has had an appendicectomy performed earlier in the day. You write a prescription for strong painkillers. The staff nurse challenges your decision and refuses to give the medication to the patient.

Choose the THREE most appropriate actions to take in this situation

A. Instruct the nurse to give the medication to the patient.
B. Discuss with the nurse why she disagrees with the prescription.
C. Ask a senior colleague for advice.
D. Complete a clinical incident form.
E. Cancel the prescription on the nurse's advice.
F. Arrange to speak to the nurse later to discuss your working relationship.
G. Write in the medical notes that the nurse has declined to give the medication.
H. Review the case again.

Ranking response format

You are looking after Mr Kucera who has previously been treated for prostate cancer. Preliminary investigations are strongly suggestive of a recurrence. As you finish taking blood from a neighbouring patient, Mr Kucera leans across and says 'tell me honestly, is my cancer back?'

Rank in order the appropriateness of the following actions in response to this situation (1 = Most appropriate; 5 = Least appropriate).

A. Explain to Mr Kucera that it is likely that his cancer has come back.
B. Reassure Mr Kucera that he will be fine.
C. Explain to Mr Kucera that you do not have all the test results, but you will speak to him as soon as you do.
D. Inform Mr Kucera that you will chase up the results of his tests and ask one of your senior colleagues to discuss them with him.
E. Invite Mr Kucera to join you and a senior nurse in a quiet room, get a colleague to hold your 'bleep' then explore his fears.

Best single response format

Patient: *So, this physiotherapy is really going to help me?*
Physician: *Absolutely, even though the first days it might still be painful.*
Patient: *Yes, I suppose it will take a while before it starts working.*
Physician: *That is why I am going to prescribe a painkiller. You should take 3 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 physician) to react to this patient's refusal to take the prescribed medication?

A. Ask her if she knows something else to relieve the pain.
B. Give her the scientific evidence as to why painkillers will help.
C. Agree not to take them now, but also stress the importance of the physiotherapy
D. Tell her that, in her own interests, she will have to start changing her attitude.

investigated the possible utility of a pictorial SJT to measure affect (Sharma & Nagle 2015), although research regarding the validity and reliability of this format of SJT delivery is in its infancy, it offers a fruitful avenue for future research.

Are SJTs fair?

Fairness issues in selection for medical education and training represents a challenge globally, and it is becoming an increasingly important consideration (Patterson et al. 2016). Regarding widening access, medical education and training providers face the challenge of identifying selection methods that both select individuals with the required skills, attributes and abilities, and also admit a diverse pool of individuals into healthcare education and training positions, to ensure that the healthcare workforce is representative of society (BMA 2009; Patterson et al. 2012c; Lievens 2014). Research allows only tentative conclusions to be drawn about the relative costs and benefits of different selection methods regarding their impact on widening access (O'Neill et al. 2013; Patterson et al. 2016), however emerging evidence suggests that SJTs may not follow socio-economic trends often observed in aptitude test scores and measures of academic attainment (Whetzel et al. 2008; Woolf et al. 2011; Wakeford et al. 2015) (i.e. in the UK, White candidates from the higher socioeconomic classes tend to outperform BMEs and those from lower socioeconomic backgrounds, and males tend to perform better than females on aptitude tests and measures of academic attainment).

A meta-analysis by Whetzel et al. (2008) reported that White test takers did marginally better than other ethnic groups, with a small effect size. Regarding gender, the research generally suggest that females tend to score marginally higher on SJTs than males (O'Connell et al. 2007; Whetzel et al. 2008; Lievens 2013; Luschin-Ebengreuth et al. 2015), which is consistent with other forms of non-academic assessment (Anderson et al. 2006).

In summary, research suggests that SJTs have less adverse impact regarding ethnicity and gender compared to other selection tools (such as cognitive ability tests). Recent research also suggests that SJTs can promote widening access compared to indicators of academic attainment.

How do candidates react when sitting SJTs?

It is important to consider candidates' reactions to selection tools because negative experiences can result in the loss of good candidates from the selection process and increase the likelihood of legal challenge (Chambers 2002; Hülsheger & Anderson 2009). Consequently, selection processes require on-going evaluation and monitoring regarding candidates' perceptions (Cascio & Aguinis 2008). Research shows fair selection processes positively influence an organisation's continued ability to attract the best candidates and recruit effectively within a given job market (Schmitt & Chan 1999). Chan and Schmitt (1997) found evidence to suggest that *face validity* perceptions (the extent to which candidates perceive a test to be relevant to the role they are applying) to significantly influence test-taking motivation. It is important therefore that candidates perceive selection methods to be relevant, face valid and fair (Gray 1997).

Focusing on SJTs more specifically, research shows that candidates have a preference for job-relevant selection methods (Gilliland 1993; Bauer et al. 2001), with work sample simulations used in selection receiving the most positive ratings from candidates alongside interviews (Hausknecht et al. 2004). Research evidence consistently shows that SJTs are positively perceived by candidates (Chan & Schmitt 2002; Lievens & Sackett 2007; Lievens et al. 2008;

Plint & Patterson 2010; Patterson et al. 2011; Koczwara et al. 2012; Roberts et al. 2014; Luschin-Ebengreuth et al. 2015), due to the highly role-specific, contextualised scenarios presented in SJTs (i.e. they are more *face valid* to candidates).

Video-based SJTs represent a medium fidelity SJT format (compared to text based SJTs which are low fidelity assessment modalities). As expected, video-based SJTs tend to have more favourable candidate reactions than text based SJTs, as candidates perceive them to have higher face validity (Chan & Schmitt 1997). However, Hausknecht et al. (2004) and Lievens and Sackett (2006) found no difference in candidate reactions to SJTs in healthcare settings, based on their method of delivery. In a recent study in which SJTs were used as an in-training assessment for junior doctors, results indicated that candidates felt that a video as an SJT delivery method was a more engaging way to view the training material compared to text based SJTs (Kerrin et al. 2014), although the SJT was used more for developmental purposes rather than selection in this context.

In summary, candidate reactions in any selection context are an important consideration to ensure that perceptions of the organisation remain favourable, and especially when considering attraction in recruitment. Candidate reactions towards SJTs used in selection for medical education and training have been found to be favourable. Although the face validity of video-based SJTs is high, they are significantly more expensive to develop and maintain compared to text-based SJTs. Research does not currently support the supposition that high-fidelity video-based SJTs would be more favourably received than text based SJTs in medical selection, although further research is required to explore the issues.

Are SJTs susceptible to coaching?

The high-stakes nature of selection into medical education and training internationally demands consideration of whether candidates' performance on SJTs can be significantly influenced by access to coaching. In overview, although early research evidence has offered mixed findings, there is now a consensus emerging.

Cullen et al. (2006) examined the coaching susceptibility of two SJTs (the College Student Questionnaire; CSQ; and the Situational Judgement Inventory; SJI). Coaching courses were developed for the two SJTs and findings showed that scores increased for the CSQ but decreased for the SJI, compared to un-coached controls. The study authors concluded that SJTs constructed from SMEs judgements are less susceptible to coaching effects. Similarly, Lievens et al. (2012) found that coaching effects were estimated to be 0.5 standard deviation difference in favour of coached candidates for an SJT for admission to medical school in Belgium.

Other researchers however have identified that knowledge based response formats are significantly less susceptible to coaching than behavioural tendency formats (McDaniel et al. 2007; Patterson et al. 2013a). Moreover, the literature suggests that building complexity into SJT scenarios can reduce susceptibility to coaching effects, as this requires candidates to engage with the scenario rather than employing a simple response strategy (Patterson et al. 2013a). Research suggests that it is possible to reduce coaching effects using an item banking approach which creates repository of test items that belong to the SJT, as well as having multiple test forms which are equated for difficulty (Patterson et al. 2013a). Many of these approaches have been used in selection for medical education and training to reduce susceptibility to coaching (Koczwara et al. 2012). A recent study concludes that coaching and revising for an SJT to select junior doctors in the UK has no effect on the operational validity the SJTs (Simon et al. 2015).

In summary, despite researchers raising concerns regarding SJTs' potential susceptibility to coaching effects, several design strategies may be adopted to minimise the potential effects of coaching where the operational validity of SJTs remains unaffected. Other research has identified that commercial coaching techniques are not as effective as previously thought (Stemig et al. 2015).

SJT in practice – How are SJTs designed and developed?

Designing an SJT requires a thorough development and validation process in line with best practice. The key steps to the best practice design of an SJT are described below. Each step ensures an SJT is relevant to the specific role for which it has been developed, and optimises the likelihood that an SJT is reliable, valid and fair.

Step 1: Role analysis and test specification

The first step in designing an SJT involves conducting a role analysis and determining the test specification. A role (job) analysis is a systematic way of collecting and analysing role-related information: typical role analysis information includes responsibilities, tasks, knowledge, skills and abilities relevant to any given role, including work-based roles and positions on educational or training courses. Not only does a role analysis contribute crucial information required for the specification of a particular selection method, such as an SJT, but it also increases the likelihood that the SJT is an accurate measurement tool.

In designing an SJT, the role analysis often includes conducting interviews with current role incumbents and collecting 'critical incidents' (challenging or salient situations that are likely to occur in the target role), which are later used to develop the item content in Step 2. By conducting a role analysis in this first step, it is possible to ensure that the SJT content is relevant to the particular role and increase fairness across candidates (Lievens et al. 2008) (Box 4).

Following the role analysis, the information collected is used to determine the *test specification*. This includes a description of the actual test content of the SJT, the types of items and response instructions, response instructions, formats used (e.g. knowledge based/behavioural tendency; multiple choice/rank order/rating/best and worst/pencil and paper; online/video-based, respectively), a description of the length of the test, the scoring convention to be used and how the test will be administered. Significant expertise is required in SJT design to design the test specification.

Box 4. Case study: Designing an SJT to target specific attributes in postgraduate training.

Each year approximately 8000 final year medical students apply for posts as junior doctors in the UK's foundation programme (Patterson et al. 2013a). Medical graduates are required to complete this two-year programme if they wish to work as doctors in the UK, and competition has increased due to the expansion of UK medical schools and the increasing number of applications from overseas.

An SJT was recently designed to be used as part of the selection process for these medical students. In order to define the professional attributes required to be successful in the Foundation Year 1 (FY1) role, a role analysis of the doctor role was performed with a person specification based on this analysis. Educational supervisors, clinical supervisors and other subject matter experts involved in the Foundation Programme contributed to the development of new test questions based on the test specification.

The results of the job analysis indicated that five professional attributes should be targeted by the SJT. These are:

- Commitment to Professionalism
- Coping with Pressure
- Effective Communication
- Patient Focus
- Working Effectively as Part of a Team

The list below shows each of the target attribute domains and includes possible SJT scenarios associated with them.

Matrix of SJT target attribute domains (Patterson et al. 2013a)

Commitment to professionalism	• Dealing with issues of confidentiality, e.g. hearing a colleague talking about a patient outside of work • Challenging inappropriate behaviour, e.g. consultant speaking to a colleague/patient in an inappropriate way • Commitment to learning, e.g. need to go to teaching while also being needed on the ward
Coping with pressure	• Knowing how to respond when you make a mistake, e.g. providing wrong medication to patient • Dealing with confrontation, e.g. with an angry relative • Seeking help when not sure of the correct procedure/best way of doing things
Effective communication	• Gathering information and communicating your intentions to nursing staff or other colleagues • Negotiating, e.g. for a scan from radiology • Listening and effectively communicating, e.g. with an unhappy patient or relative
Patient focus	• Identifying that a patient's views and concerns are important and they should have input into their care • Considering that a patient may have different needs from others around them • Spending time trying to understand a patient's concerns and empathising with them
Working effectively as part of a team	• Recognising and valuing the skills and knowledge of nursing staff, when faced with a disagreement about a patient's care • Consulting with colleagues about how to share workload fairly • Offering assistance and support to colleagues when they are not able to handle their workload

Step 2: Item development and initial reviews

Having documented a test specification and collected role analysis information, SJT scenarios and response options are developed in collaboration with individuals who are familiar with the target role (SMEs). These may include role incumbents and supervisors, or any other staff familiar with the target role. Working with these SMEs is essential for item development in order to ensure that SJT scenarios and responses are developed based on realistic, appropriate and plausible scenarios. The 'critical incidents' identified in Step 1 can be used to develop different role-related scenarios where the candidate would need to make a decision regarding the best course of action, given the situation. Different responses to these scenarios are also developed in conjunction with SMEs, who advise the item writer regarding the appropriateness or importance of various response options (i.e. what the scoring key should be for each response). A thorough and systematic review of these scenarios and responses is then undertaken by a different set of SMEs to ensure that each item is fair, relevant to the role and realistic.

Step 3: Scoring key agreed by subject matter experts

Once item content has been developed, agreement between SMEs is required on the scoring key (i.e. how the response should be scored) for each of the possible responses to the given scenario. This is typically achieved through a *concordance panel* with SMEs. Generally this conducted with a different group of SMEs to those during Step 2, who score each response option based on how appropriate or important they consider each response to be. This process is used for the various response options available in an SJT, including ranking responses from most to least appropriate, choosing the single best answer or the three best answers (multiple choice), or rating each response independently of each other. This is often an iterative process, but results in a final scoring key being developed.

Step 4: Test construction

In Step 4, the test is constructed. It may be in a written (paper and pencil, or electronic) format, or in some situations, the scenario may be presented in a video or interactive format.

Step 5: Piloting

Once the SJT has been constructed, the next step is piloting to ensure that it is fair and measures what it is intended to measure (i.e. has construct validity). Piloting also provides an opportunity to gain candidate reactions to the SJT, for example whether candidates perceive it to be fair and relevant to the target role. This minimises the risk of potential legal action against the organisation in operational selection settings due to a perceived lack of fairness or robustness (Patterson et al. 2011).

Step 6: Psychometric analysis and quality assurance

After an SJT has been piloted, psychometric analysis of the pilot data can be conducted. At this stage, it is possible to examine the reliability and validity of the test, and to ensure

that each SJT item is performing well psychometrically. In addition, a fairness analysis can be conducted to identify whether there are any performance differences on the SJT as a whole, and for specific scenarios, based on demographic group differences such as ethnicity or gender. If performance differences do exist, it is possible that an item is discriminating against a particular sub-group (differential item functioning) and consideration should be given to whether it is fair and appropriate for the scenario to remain in the test.

Step 7: Development of an item bank

Finally, there will be further development of the item bank through ongoing development, review and validation of the scenarios and responses as more candidate score data is collected.

Approaches to scoring SJTs

There are various scoring methods for multiple choice SJTs. They are typically broken down into *rational* and *empirical*. Rational scoring of items is based on experts' judgement concerning the effectiveness of responses or best/worst options. When items are empirically scored, they are administered to a large pilot sample. Items are selected and/or weighted according to evidence that they differentiate between individuals who score at different levels on a criterion variable (e.g. job performance).

Conclusions and future directions for research and practice

SJTs represent a measurement methodology, rather than a specific measure *per se*, as each SJT may be constructed differently; and so have the potential to vary significantly in terms of their robustness, reliability and validity. Crucially therefore, evidence suggests that good quality SJTs require significant expertise to design and develop appropriately. Best practice design of SJTs begins with a thorough role analysis, constructed in consultation with SMEs. Design issues should also consider susceptibility to coaching, especially in selection for medical education and training.

On balance, the evidence to date demonstrates that SJTs are a valid, reliable and well-received method for measuring important non-academic attributes, such as empathy, integrity and teamwork. SJTs have been successfully implemented for use in selection across a range of healthcare professions, especially within medical education and training. SJTs have the benefit of having reduced sub-group differences compared to other selection methods, and are well received by candidates. The theoretical basis underpinning SJTs shows they assess individuals' beliefs about the costs and benefits of expressing certain traits via behaviours in given situations. Future research could extend the emerging evidence relating to the construct validity of SJTs, for example, exploring why there is little effect of socio-economic status and SJT performance (unlike indicators of academic attainment).

Despite a plethora of research evidence supporting the use of SJTs in the context of medical education and training, a number of possible areas of exploration remain. Further research could systematically compare video-based and text based SJTs to more adequately explain the advantages and disadvantages of each approach, where potentially researchers are able to comment on what each mode of item presentation offers more fully.

Future areas of enquiry may include the use of SJTs as a diagnostic tool to identify individuals' training needs. This has the potential to be supplemented by education interventions to provide training in the non-academic skills required for the target role, and so accelerate time to competence in preparation for entry into clinical practice. Although more traditionally used in selection contexts, extending the application of SJTs to training and development may provide a beneficial resource in monitoring and training relevant non-academic attributes as individuals progress through a given role or training programme.

Further longitudinal research studies are required to evaluate the extent to which SJTs effectively predict performance throughout the medical education pathway, from medical school admissions through to independent clinical practice, and beyond. This is especially relevant given the emerging evidence that SJTs have different predictive validity at different stages during medical education, training and practice. More research is also required to confirm initial findings that SJTs are more predictive of performance at the lower end of score distributions than at the top end (i.e. in identifying individuals who lack the appropriate values and attributes to be suited to a career in healthcare). This, in turn, would help researchers develop the theory underpinning SJTs as a measurement method.

To conclude, SJTs represent a reliable, valid, well-received and fair selection method when designed appropriately. Future research should investigate the use of SJTs for development, focus on gathering longitudinal data, and assess the differential predictive validity of SJTs at different points during education and training pathways in healthcare and at different ends of score distributions.

Notes on contributors

PROFESSOR FIONA PATTERSON, BSc, MSc, PhD, CPsychol, AcSS, FRSA, FCMI, FRCGP (Hon), is a leading expert in the field of selection, assessment and innovation in organisations with over 20 years' experience working with organisations at a strategic level. Fiona holds a Principal Researcher post at the University of Cambridge and a Visiting Chair at City University, London. Over the past 15 years, her research has had a major impact upon governmental and corporate policy in the UK and abroad.

DR LARA ZIBARRAS, BSc, MSc, PhD, C.Psychol, is Senior Lecturer in Occupational Psychology at City University, London. Her key research areas are employee selection, particularly the candidate's perspective and new selection methodologies such as situational judgement tests and pro-environmental behaviour in the workplace. She has published widely in academic journals and presented her research at both national and international conferences.

VICKI ASHWORTH, BSc, MSc, CPsychol, is an Associate Director at Work Psychology Group. Vicki specialises in the fields of assessment and selection, evaluation and organisational development, delivering bespoke solutions to public and private sector clients in the UK and internationally.

Acknowledgements

We especially would like to thank Anna Rosselli and Fran Cousans for their support in preparing the final manuscript.

Declaration of interest: FP and VA provide advice to Health Education England on selection methodology. However, they do not receive royalties for any methodology used. The authors alone are responsible for the content and writing of the article.

References

Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. 2003. Assessing personal qualities in medical school admissions. Acad Med 78(3):313–321.

Anderson N, Lievens F, van Dam K, Born M. 2006. A construct-driven investigation of gender differences in a leadership-role assessment center. J Appl Psychol 91(3):555–566.

Arnold J, Randall R, Patterson F, Silvester J, Robertson I, Cooper C, Burnes B, Swailes S, Harris D, Axtell C, Den Hartog D. 2010. Work psychology: Understanding human behaviour in the workplace. 5th ed. England: Pearson Education Limited.

Ashworth V, Fung K, Shaw R. 2014. National University of Singapore Situational Judgment Test and Focused Skills Assessment: Analysis & Evaluation.

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

Birkeland SA, Manson TM, Kisamore JL, Brannick MT, Smith MA. 2006. A meta-analytic investigation of job applicant faking on personality measures. Int J Sel Assess 14(4):317–335.

BMA. 2009. Equality and diversity in UK medical schools. London: BMA.

Campbell J, Gasser M, Oswald F. 1996. The substantive nature of job performance variability. In: Murphy KR, editor. Individual differences and behavior in organizations. San Francisco: Jossey-Boss. pp 258–299.

Cascio WF, Aguinis H. 2008. Staffing twenty-first-century organizations. Acad Manag Ann 2(1):133–165.

Catano VM, Brochu A, Lamerson CD. 2012. Assessing the reliability of situational judgment tests used in high-stakes situations. Int J Sel Assess 20(3):333–346.

Cavendish C. 2013. The Cavendish Review: An Independent Review into Healthcare Assistants and Support Workers in the NHS and social care settings. England: Department of Health.

Chambers BA. 2002. Applicant reactions and their consequences: Review, advice, and recommendations for future research. Int J Manag Rev 4(4):317–333.

Chan D, Schmitt N. 1997. Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. J Appl Psychol 82(1): 143–159.

Chan D, Schmitt N. 2002. Situational judgment and job performance. Hum Perform 15(3):233–254.

Chan D, Schmitt N, Sacco JM, DeShon RP. 1998. Understanding pretest and posttest reactions to cognitive ability and personality tests. J Appl Psychol 83(3):471–485.

Christian MS, Edwards BD, Bradley JC. 2010. Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. Pers Psychol 63(1):83–117.

Clevenger J, Pereira GM, Wiechmann D, Schmitt N, Harvey VS. 2001. Incremental validity of situational judgment tests. J Appl Psychol 86(3):410–417.

Crook AE, Beier ME, Cox CB, Kell HJ, Hanks AR, Motowidlo SJ. 2011. Measuring relationships between personality, knowledge, and performance using single-response situational judgment tests. Int J Sel Assess 19(4):363–373.

Cullen MJ, Sackett PR, Lievens F. 2006. Threats to the operational use of situational judgment tests in the college admission process. Int J Sel Assess 14(2):142–155.

Dore KL, Reiter HI, Eva KW, Krueger S, Scriven E, Siu E, Hilsden S, Thomas J, Norman GR. 2009. Extending the interview to all medical school candidates – Computer-based multiple sample evaluation of noncognitive skills (CMSENS). Acad Med 84(10):S9–S12.

Escudier M, Flaxman C, Cousans F, Woolford M, Patterson F. 2015. Pilot results from a new assessment of professionalism. Proceedings of the AMEE Conference 2015, Glasgow, UK.

Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. 2009. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ 43(8):767–775.

Ferguson E, James D, Madeley L. 2002. Factors associated with success in medical school: Systematic review of the literature. BMJ 324(April): 952–957.

Ferguson E, James D, O'Hehir F, Sanders A. 2003. Pilot study of the roles of personality, references, and personal statements in relation to performance over the five years of a medical degree. BMJ. 326:429–432.

Ferguson E, Sanders A, O'Hehir F, James D. 2000. Predictive validity of personal statements and the role of the five-factor model of personality in relation to medical training. J Occup Organ Psychol 73(3):321–344.

Ferguson E, Semper H, Yates J, Fitzgerald JE, Skatova A, James D. 2014. The "dark side" and "bright side" of personality: When too much conscientiousness and too little anxiety are detrimental with respect to the acquisition of medical knowledge and skill. PLoS One 9(2):1–11.

Francis R. 2013. Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry: Executive Summary. England: NHS Foundation Trust.

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

Gilliland SW. 1993. The perceived fairness of selection systems: An organizational justice perspective. Acad Manag Rev 18(4):694–734.

Gray MJA. 1997. Evidence-based healthcare. London: Churchill Livingstone.

Hausknecht JP, Day DV, Thomas SC. 2004. Applicant reactions to selection procedures: An updated model and meta-analysis. Pers Psychol 57(3):639–683.

Hülsheger UR, Anderson N. 2009. Applicant perspectives in selection: Going beyond preference reactions. Int J Sel Assess 17(4):335–345.

Husbands A, Mathieson A, Dowell J, Cleland JA, MacKenzie R. 2014. Predictive validity of the UK clinical aptitude test in the final years of medical school: A prospective cohort study. BMC Med Educ 14(1):88.

James D, Yates J, Nicholson S. 2010. Comparison of A level and UKCAT performance in students applying to UK medical and dental schools in 2006: Cohort study. BMJ 340:c478.

Kerrin M, Aitkenhead A, Shaw R. 2014. Public Health Situational Judgement Test 2014: Final Report.

Kerrin M, Rowett E, Lopes S. 2015. The University of Nottingham: School of Veterinary Medicine & Science (SVMS): Situational Judgement Test Pilot & Operational Delivery 2014.

Kline P. 2000. Handbook of Psychological Testing, Vol. 12, 2nd ed. London, UK: Routledge.

Koczwara A, Ashworth V. 2013. Selection and assessment. In: Lewis R, Zibarras L, editors. Work and occupational psychology: Integrating theory and practice. London: SAGE. pp 295–342.

Koczwara A, Patterson F, Zibarras L, Kerrin M, Irish B, Wilkinson M. 2012. Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. Med Educ 46(4):399–408.

Kreiter C, Axelson R. 2013. A perspective on medical school admission research and practice over the last 25 years. Teach Learn Med 25(S1): S50–S56.

Libbrecht N, Lievens F, Carette B, Côté S. 2014. Emotional intelligence predicts success in medical school. Emotion 14(1):64–73.

Lievens F. 2013. Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests. Med Educ 47:182–189.

Lievens F. 2014. Diversity in medical school admission: Insights from personnel recruitment and selection. Med Educ 49(1):7–20.

Lievens F, Buyse T, Sackett PR. 2005a. The operational validity of a video-based situational judgment test for medical college admissions:

Illustrating the importance of matching predictor and criterion construct domains. J Appl Psychol 90(3):442–452.
Lievens F, Buyse T, Sackett PR. 2005b. Retest effects in operational selection settings: Development and test of a framework. Pers Psychol 58(4):981–1007.
Lievens F, Buyse T, Sackett PR, Connelly BS. 2012. The effects of coaching on situational judgment tests in high-stakes selection. Int J Sel Assess 20(3):272–282.
Lievens F, Patterson F. 2011. 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 96(5):927–940.
Lievens F, Peeters H, Schollaert E. 2008. Situational judgment tests: A review of recent research. Pers Rev 37(4):426–441.
Lievens F, Sackett PR. 2006. Video-based versus written situational judgment tests: A comparison in terms of predictive validity. J Appl Psychol 91(5):1181–1188.
Lievens F, Sackett PR. 2007. Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. J Appl Psychol 92(4):1043–1055.
Lievens F, Sackett PR. 2012. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. J Appl Psychol 97(2):460–468.
Lopes S, Baron H, Patterson F. 2015a. Specialty recruitment assessment results & scoring: Psychiatry CT1 entry. Report to HEE.
Lopes S, Baron H, Patterson F. 2015b. Specialty recruitment assessment results & scoring: Ophthalmology ST1 entry. Report to HEE.
Luschin-Ebengreuth M, Dimai HP, Ithaler D, Neges HM, Reibnegger G. 2015. Situational judgment test as an additional tool in a medical admission test: An observational investigation. BMC Res Notes 8(1):1–7.
Martin M, Motowidlo SJ. 2010. A single-response SJT for measuring procedural knowledge for human factors professionals. Poster Session Presented at 25th Annual Meeting of the Society for Industrial Organizational Psychology, Atlanta, GA.
McCrae RR, Costa PT. 2008. Empirical and Theoretical status of the five-factor model of personality traits. In: Boyle GJ, Matthews G, Saklofske DH, editors. SAGE handbook of personality theory and assessment. London: SAGE Publications. pp 273–294.
McDaniel MA, Hartman NS, Grubb, III WL. 2003. Situational judgment tests, knowledge, behavioral tendency and validity: A meta-analysis. 18th Annual Conference of the Society for Industrial Organizational Psychology Orlando.
McDaniel MA, Hartman NS, Whetzel D, Grubb III WL. 2007. Situational judgment tests, response instructions, and validity: A meta-analysis. Pers Psychol 60(1):63–91.
McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. 2001. Use of situational judgment tests to predict job performance: A clarification of the literature. J Appl Psychol 86(4):730–740.
McDaniel MA, Nguyen NT. 2001. Situational judgment tests: A review of practice and constructs assessed. Int J Sel Assess 9(1/2):103–113.
McDaniel MA, Whetzel D. 2007. Situational judgment tests. In: Whetzel DL, Wheaton GR, editors. Applied measurement: Industrial psychology in human resources management. Mahwah, NJ: Lawrence Erlbaum and Associates. pp 235–257.
McDaniel MA, Whetzel D, Schmidt FL, Maurer SD. 1994. The validity of employment interviews: A comprehensive review and meta-analysis. J Appl Psychol 79(4):599–616.
McManus IC, Dewberry C, Nicholson S, Dowell J. 2013. The UKCAT-12 study: Educational attainment, aptitude test performance, demographic and socio-economic contextual factors as predictors of first year outcome in a cross-sectional collaborative study of 12 UK medical schools. BMC Med 11:244.
McManus IC, Woolf K, Dacre J. 2008. Even one star at A level could be "too little, too late" for medical student selection. BMC Med Educ 8: 1–4.
Motowidlo SJ. 2003. Handbook of psychology. Hoboken, NJ: John Wiley & Sons, Inc.
Motowidlo SJ, Beier ME. 2010. Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. J Appl Psychol 95(2):321–333.
Motowidlo SJ, Crook AE, Kell HJ, Naemi B. 2009. Measuring procedural knowledge more simply with a single-response situational judgment test. J Bus Psychol 24(3):281–288.
Motowidlo SJ, Dunnette MD, Carter GW. 1990. An alternative selection procedure: The low-fidelity simulation. J Appl Psychol 75(6): 640–647.
Motowidlo SJ, Hooper AC, Jackson HL. 2006. Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. J Appl Psychol 91(4):749–761.
Mullins M, Schmitt N. 1998. Situational judgment testing: Will the real constructs please present themselves? 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX..
Nguyen NT, SBiderman MD, SMcDaniel MA. 2005. Effects of response instructions on faking a situational judgment test. Int J Sel Assess 13(4):250–260.
NHS. 2013. NHS Constitution. Englan: NHS.
O'Connell MS, Hartman NS, McDaniel MA, Grubb III WL, Lawrence A. 2007. Incremental validity of situational judgment tests for task and contextual job performance. Int J Sel Assess 15(1):19–29.
O'Neill L, Vonsild MC, Wallstedt B, Dornan T. 2013. Admission criteria and diversity in medical school. Med Educ 47(6):557–561.
Oswald F, Schmitt N, Kim B, Ramsay LJ, Gillespie MA. 2004. Developing a biodata measure and situational judgment inventory as predictors of college student performance. J Appl Psychol 89(2):187–207.
Parks L, Guay RP. 2009. Personality, values, and motivation. Pers Individ Dif 47(7):675–684.
Patterson F, Aitkenhead A, Edwards H, Flaxman C, Shaw R, Rosselli A. 2015a. Analysis of the situational judgement test for selection to the foundation programme 2015: Technical Report.
Patterson F, Ashworth V, Kerrin M, O'Neill P. 2013a. Situational judgement tests represent a measurement method and can be designed to minimise coaching effects. Med Educ 47(2):220–221.
Patterson F, Ashworth V, Mehra S, Falcon H. 2012a. Could situational judgement tests be used for selection into dental foundation training? Br Dent J 213(1):23–26.
Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. 2012b. Evaluations of situational judgement tests to assess non-academic attributes in selection. Med Educ 46:850–868.
Patterson F, Baron H, Carr V, Plint S, Lane P. 2009a. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. Med Educ 43(1):50–57.
Patterson F, Carr V, Zibarras L, Burr B, Berkin L, Plint S, Irish B, Gregory S. 2009b. New machine-marked tests for selection into core medical training: Evidence from two validation studies. Clin Med (Northfield, IL) 9(5):417–420.
Patterson F, Ferguson E, Lane P, Farrell K, Martlew J, Wells A. 2000. A competency model for general practice: Implications for selection, training, and development. Br J Gen Pract 50(452):188–193.
Patterson F, Ferguson E, Thomas S. 2008. Using job analysis to identify core and specific competencies: Implications for selection and recruitment. Med Educ 42(12):1195–1204.
Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland JA. 2016. How effective are selection methods in medical education and training? Evidence from a systematic review. Med Educ [in press].
Patterson F, Lievens F, Kerrin M, Munro N, Irish B. 2013b. The predictive validity of selection for entry into postgraduate training in general practice: Evidence from three longitudinal studies. Br J Gen Pract 63(616):734–741.
Patterson F, Lievens F, Kerrin M, Zibarras L, Carette B. 2012c. Designing selection systems for medicine: The importance of balancing predictive and political validity in high-stakes selection contexts. Int J Sel Assess 20(4):486–496.
Patterson F, Martin S. 2014. UKCAT SJT: A study to explore validation methodology and early findings.
Patterson F, Martin S, Baron H, Fung K, Flaxman C. 2014. UKCAT situational judgement test: Technical report.
Patterson F, Prescott-Clements L, Zibarras L, Edwards H, Kerrin M, Cousans F. 2015b. Recruiting for values in healthcare: A preliminary review of the evidence. Adv Heal Sci Educ 1–23.

Patterson F, Zibarras L, Carr V, Irish B, Gregory S. 2011. Evaluating candidate reactions to selection practices using organisational justice theory. Med Educ 45(3):289–297.

Plint S, Patterson F. 2010. Identifying critical success factors for designing selection processes into postgraduate specialty training: The case of UK general practice. Postgrad Med J 86(1016):323–327.

Ployhart RE, Ehrhart MG. 2003. Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. Int J Sel Assess 11(1):1–16.

Ployhart RE, Weekley J, Holtz BC, Kemp C. 2003. Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? Pers Psychol 56(3):733–752.

Poole P, Moriarty HJ, Wearn AM, Wilkinson T, Weller JM. 2009. Medical student selection in New Zealand: Looking to the future. N Z Med J 122(1306):88–100.

Prideaux D, Roberts C, Eva KW, Centeno A, McCrorie P, McManus IC, Patterson F, Powis D, Tekian A, Wilkinson D. 2011. Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach 33(3):215–223.

Rankin B. 2013. Emotional intelligence: Enhancing values-based practice and compassionate care in nursing. J Adv Nurs 69(12):2717–2725.

Roberts C, Clark T, Burgess A, Frommer M, Grant M, Mossman K. 2014. The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training. BMC Med Educ 14:1–11.

Roberts C, Togno JM. 2011. Selection into specialist training programs: An approach from general practice. Med J Aust 194(2):93–95.

Roth PL, Bobko P, Buster MA. 2013. Situational judgment tests: The influence and importance of applicant status and targeted constructs on estimates of Black-White subgroup differences. J Occup Organ Psychol 86(3):394–409.

Rust J, Golombok S. 1999. Modern psychometrics: The science of psychological assessment, Vol. 2. London: Routledge.

Sartania N, McClure JD, Sweeting H, Browitt A. 2014. Predictive power of UKCAT and other pre-admission measures for performance in a medical school in Glasgow: A cohort study. BMC Med Educ 14(1):116–125.

Schmidt FL. 1994. The future of personnel selection in the U.S. Army. In Weekley JA, Ployhart RE, editors. Personnel selection and classification. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. pp 333–350.

Schmidt FL, Hunter JE. 1998. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychol Bull 124(2):262–274.

Schmitt N, Chan D. 1999. The status of research on applicant reactions to selection tests and its implications for managers. Int J Manag Rev 1(1):45–62.

Sharma N, Nagle YK. 2015. Development of pictorial situational judgement test of affect. Psychology 6(March):400–408.

Simon E, Walsh K, Paterson-Brown F, Cahill D. 2015. Does a high ranking mean success in the situational judgement test? Clin Teach 12(1):42–45.

Stemig MS, Sackett PR, Lievens F. 2015. Effects of organizationally endorsed coaching on performance and validity of situational judgment tests. Int J Sel Assess 23(2):174–181.

St-Sauveur C, Girouard S, Goyette V. 2014. Use of situational judgment tests in personnel selection: Are the different methods for scoring the response options equivalent? Int J Sel Assess 22(3):225–239.

Wakeford R, Denney M-L, Ludka-Stempien K, Dacre J, McManus IC. 2015. Cross-comparison of MRCGP & MRCP(UK) in a database linkage study of 2,284 candidates taking both examinations: Assessment of validity and differential performance by ethnicity. BMC Med Educ 15(1):1–12.

Weekley J, Ployhart RE. 2005. Situational judgment: Antecedents and relationships with performance. Hum Perform 18(1):81–104.

Weekley J, Ployhart RE, Holtz BC. 2006. On the development of situational judgment tests: Issues in item development, scaling, and scoring. In: Weekley JA, Ployhart RE, editors. Situational judgment tests: Theory, measurement, and application. Mahwah, NJ: Lawrence Erlbaum Associates, Inc 157–182.

Wernimont P, Campbell J. 1968. Signs, samples, and criteria. J Appl Psychol 52(5):372–376.

Whetzel D, McDaniel MA. 2009. Situational judgment tests: An overview of current research. Hum Resour Manag Rev 19(3):188–202.

Whetzel D, McDaniel MA, Nguyen NT. 2008. Subgroup differences in situational judgment test performance: A meta-analysis. Hum Perform 21:291–309.

Woolf K, Potts HWW, McManus IC. 2011. Ethnicity and academic performance in UK trained doctors and medical students: Systematic review and meta-analysis. BMJ 342:d901.

Wyatt MRR, Pathak SB, Zibarras L. 2010. Advancing selection in an SME: Is best practice methodology applicable? Int Small Bus J 28(3):258–273.