Article:
Tiffin, Paul Alexander orcid.org/0000-0003-1770-5034, Paton, Lewis William orcid.org/0000-0002-3328-5634, O'Mara, Deborah et al. (3 more authors) (2020) Situational judgments tests for selection:traditional vs construct-driven approaches. Medical Education. pp. 105-115. ISSN 0308-0110

https://doi.org/10.1111/medu.14011

The cross-cutting edge: situational judgments tests for selection: traditional versus construct-driven approaches

Paul A. Tiffin, Lewis W. Paton, Deborah O'Mara, Carolyn MacCann, Jonas W.B. Lang and Filip Lievens

Author details:

PAT: Department of Health Sciences, University of York and Health Professions Education Unit, Hull York Medical School, UK, paul.tiffin@york.ac.uk

LWP: Department of Health Sciences, University of York, UK, lewis.paton@york.ac.uk

DOM: University of Sydney Medical School, Faculty of Medicine & Health, University of Sydney, Australia, deborah.omara@sydney.edu.au

CM: School of Psychology, University of Sydney, Australia, carolyn.maccann@sydney.edu.au

JL: Department of Personnel Management, Work and Organizational Psychology, Ghent University, Belgium, & the Business School, University of Exeter, Jonas.Lang@ugent.be

FL: Lee Kong Chian School of Business, Singapore Management University, Singapore, filiplievens@smu.edu.sg

Abstract

Context

Historically, Situational Judgement Tests (SJTs) have been widely used for personnel selection. Their use in medical selection in Europe is growing with plans for further expansion into North America and Australasia in an attempt to measure and select on 'non-academic' personal attributes. However, there is a lack of clarity regarding what such tests actually measure and how they should be designed, scored and implemented within the medical and health education selection process. In particular, the theoretical basis from which such tests are developed will determine the scoring options available, influencing their psychometric properties and, ultimately, their validity.

Objective

The aim of this article is to create an awareness of the previous theory and practice that has informed SJT development. We describe the emerging interest in the use of the SJT format to measure specific constructs (e.g. 'resilience', 'dependability' etc.), drawing on the tradition of 'individual differences' psychology. We compare and contrast this newer 'construct-driven' method with the traditional, pragmatic approach to SJT creation, often employed by organisational psychologists. Making reference to measurement theory, we highlight how the anticipated psychometric properties of traditional versus construct-driven SJTs are likely to differ.

Conclusions

Compared to traditional SJTs, construct-driven SJTs have a strong theoretical basis, are uni- rather than multidimensional, and may behave more like personality self-report instruments. Emerging evidence also suggests that construct-driven SJTs have comparable predictive validity for workplace performance, although they may be more prone to 'faking' effects. It is possible that construct-driven approaches prove more appropriate at early stages of medical selection, where candidates have little or no healthcare work experience. Conversely, traditional SJTs may be more suitable for

specialty recruitment, where a range of hypothetical workplace scenarios can be sampled in assessments.

A brief history of SJTs for personnel selection…

A situational judgement test (SJT) is an assessment whereby a candidate is presented with a specific scenario and must evaluate several possible responses to the scenario. The response format can vary but commonly involve either ranking potential behavioural responses in order of appropriateness or perceived effectiveness. Another commonly employed choice format involves a candidate choosing the 'best' and 'worst' behaviours depicted. An example of a ranking format SJT is shown in Figure 1. A frequent alternative is a rating scale type response format, which involves rating examples of behavior according to some attribute, such as 'appropriateness' as shown in Figure 2.

The SJT approach to assessment has been used in personnel selection for over half a century. Notably, SJTs were used during World War II to evaluate the judgment of soldiers. In the 1960s assessments using the SJT format were developed in an attempt to measure the leadership potential of job applicants (1). Early examples of this approach to testing include the 'Practical Judgment Test' (2). The use of SJTs in personnel selection became much more widely adopted in the 1990s. This popularity was probably triggered by the reconceptualisation of SJTs as 'low-fidelity simulations' (3). That is, the test was intended to mimic the kinds of workplace situations likely to be encountered by a successful candidate appointed to a job or training role.

SJTs and medical selection

Internationally, there is high competition for admission to medical school and some postgraduate training. This has led to a strong emphasis on academic achievement as a key selection mechanism. For example, secondary (high) school and also cognitive performance, assessed via 'aptitude tests' such as the Universities Clinical Aptitude Test (UCAT- formerly 'UKCAT'), the Graduate Medical Schools Admission Test (GAMSAT) and the Medical College Admission Test (MCAT) are frequently used to select students for medical education. However, there has been an increasing recognition that it is non-academic personal qualities (sometimes referred to as 'non-cognitive' traits) that are as least as important to effective functioning as a practising physician (4). Indeed, for the majority of UK licensed doctors who have

been censured by the regulator, it is personal rather than clinical behavior that led to the initial allegation (5). In line with the recognition that 'non-academic' qualities are also critical in medicine, SJTs are currently being used in Europe as part of medical selection, usually in order to complement more resource intensive selection methods that involve direct interviewing. The SJT format is already being used in Australasia as part of recruitment into postgraduate medical training with plans to roll out an undergraduate selection SJT as part of the introduction of the UCAT into Australia. Moreover, a pilot project is being conducted for incorporating an SJT, aiming to evaluate an applicant's understanding of 'pre-professional behaviours', as part of the MCAT used in North America (6). It should be emphasised that SJTs represent an assessment approach, not a content area, and that test format and content should not be confused (7). Thus, this paper focuses only on SJTs that are used to evaluate 'non-academic' personal qualities as part of personnel selection processes.

Despite the rapidly increasing use of the SJT format in the medical education selection process, there is no agreement on what such assessments are generally measuring or how they should be optimally designed and implemented. In this article we will first describe the traditional SJT development approach, now commonly used in medical selection. We also outline an emerging interest in construct-driven approaches to developing SJT-based assessments that can be used in personnel selection. In contrast to traditional SJT approaches, this latter movement owes much to the 'individual differences' field of psychology. This has had an emphasis on developing psychometric instruments capable of accurately discriminating between individuals depending on the level of a specifically defined trait or ability they possess. We compare and contrast the likely advantages and limitations of each of these two approaches.

In order to make informed decisions about how SJTs should be developed and used within medical selection it would seem essential to understand the conceptual bases for the two main approaches to their creation. This article will focus on contrasting aspects of development, and the resulting psychometric properties of SJT format assessments used in this context. A summary of the main differences between traditional and construct-driven SJTs is provided in Table 1.

Traditional SJTs for personnel selection

The rationale for traditional SJT development could be considered primarily pragmatic, and largely atheoretical. As aforementioned, the popularity of SJTs for personnel selection was encouraged by their reframing as 'low-fidelity simulations' (3). A simulation, in this sense, is a selection procedure intended to mimic psychological or physical aspects of the job (8). The 'low fidelity' term is taken to mean that responses do not involve actual enactment of the behaviours that would be expected in the workplace scenario depicted. Rather, responding involves selecting, from among a number of closed-ended responses, using knowledge of the appropriateness of intentions to behave in a range of ways (9). Therefore, the ability of the test to predict future job performance rests on the idea that the simulation corresponds to actual future work situations and that test-takers will exhibit behavioural consistency (10). This latter term implies that candidate responses to the test will accurately reflect their actual, future, workplace behaviours (11). In this sense traditional SJTs used in personnel selection could be considered special cases of a knowledge test. Such knowledge would include procedural knowledge about *what* to do in certain situations and *how* to do it.

In the traditional approach to SJT development the description of the workplace situations are of key importance. This is because the way the scenes are depicted will situate a respondent's perception of the scenarios. Ideally the candidate will be able to use these descriptions to imagine themselves in that context and be able to make a judgment about the behaviour they would, or should, exhibit, in that situation (12). Thus, the ultimate goal of traditional SJTs is to devise a test to sample the domain of interest, to present people with a representative set of situations in this domain, and to assume that the scores that capture how their procedural knowledge of how to respond to those situations will demonstrate criterion-related validity. That is, the scores achieved on the test will be correlated with future relative workplace performance, for example, as judged by a supervisor. That is, the aim of SJTs, developed according to this framework is to capture a test taker's contextualized responses to samples of workplace situations.

Traditional SJT development

The traditional SJT creation process typically follows a conventional path (Figure 3) which involves: 1) generating 'critical incidents' from subject matter experts (SMEs) in a particular field of work, 2) capturing possible responses to the incidents from other SMEs and novices; and 3) creating a scoring key from another group of SMEs (3, 13). Sometimes Step 1 is preceded by a 'blue printing' process involving obtaining the views of SMEs on which characteristics they deem relevant to effective workplace performance. Scenarios can then be selected that are perceived to evaluate the traits included in the blueprint. The scenarios themselves can be presented in a variety of formats; text, multimedia, 3D animated or even avatar-based. Figure 2 shows an example of text-based SJT that was created for undergraduate medical selection. SMEs were involved in appraising a provisional pilot set of items and constructing a scoring key based on consensus regarding the best response to each item. There are no current guidelines on the number of SMEs recommended for this process, but most developers use 12 to 30.

For all response formats, the score allocated to the response categories for each item is usually based on the similarity of the candidate's responses to the responses determined by SMEs. The only exception to this would be to either use a norm referenced approach (i.e. 'the wisdom of the crowd') or train a machine learning algorithm to predict an outcome from the scoring pattern (14). Thus, scores can be calculated in many ways and the scoring scheme influences the properties of the final test scores (15).

The type of instruction given to candidates of SJTs is also relevant to the validity evidence regarding cognitive processes. In general, SJTs that instruct candidates to select a response based on what they *would* do in that situation tend to have scores that correlate more highly with personality measures. In contrast, those that instruct a candidate to indicate what *should* they do tend to show lower correlations with personality measures (16). The latter type of test could be considered a special type of knowledge test, and by definition, less prone to 'faking' effects and social desirability bias (17). That is, either a candidate knows what should be done or they do not. Yet, this difference between SJT scores on the two instructions was found only in research contexts and non-high stakes contexts. In a high stakes context all candidates seem to adopt a '*should do'* response frame, regardless of the actual

instruction (18). Scoring strategy may also influence the impact of faking attempts (19).

There is meta-analytic evidence for the overall predictive validity of the scores derived from such traditional SJTs to predict future work performance (16, 20). Moreover, a previous review of the use of SJTs for the evaluation of non-academic traits in general concluded that there was evidence of cost-effectiveness compared to other approaches (e.g. interviews) (21). Emerging evidence also indicates that, to date, in a variety of undergraduate and postgraduate medical selection settings, at least modest correlations are observed between performance on these traditional SJTs and subsequent relevant outcomes that reflect aspects of social functioning (22-24). This suggests, at least in this context, such selection SJTs are generally estimating, to some extent, knowledge of inter-personal functioning. Specifically, higher performing candidates are tending to respond to the items in a way that corresponds to the choices of the SMEs used to derive a scoring key.

This rather pragmatic approach to assessment development often results in a lack of clarity over what is actually being measured. Indeed, the construct-related validity of SJTs (that is, whether they actually measure the constructs they purport to) has been described as a 'hot mess' (25). One might ask, if the scores from traditional SJTs tend to predict future candidate performance, does it matter what they are actually measuring? Firstly, there are practical challenges with deploying tests where the dimension/s (constructs) being measured are not well defined. For example, for security reasons, wide-scale testing often utilises more than one form of the test. In the interests of fairness all forms of the SJT forms should be 'equated'. That is, candidates of equal ability should obtain the same score, irrespective of the test form to which they are allocated. Such equating is extremely difficult to assure in the absence of a well-defined 'measurement model', where an observed score is tightly linked to a particular construct. However, this risk can be offset, to some extent, by careful "blue printing" of content across a relatively small number of test forms (say two or three), thus generating alternative, though similar, forms of the test with a similar mix of scenario-related material (26). Conversely, it has been highlighted that the ability to design SJTs that have an established relationship with well-defined constructs has many potential benefits (9).

Construct-driven selection SJTs

Theoretical basis

As mentioned earlier, the widespread use of SJTs for personnel selection was stimulated by framing them as low-fidelity simulations (3). However, some years later Motowidlo also provided a theoretical basis for the effectiveness of SJTs, hypothesising that, in this context, they were a way of eliciting and measuring 'implicit trait policies' (ITPs) (27). An ITP is conceptualised as a set of implicit beliefs about causal relations between personality traits and behavioural effectiveness. The proponents of this theory provide an illustrative example related to the personality trait of agreeableness. They argue that if depicted actions in the SJT response options that express high agreeableness are truly more effective than actions that express low agreeableness, more agreeable people will weigh response agreeableness more heavily. This is in contrast to less agreeable people, who will tend to perceive the same responses as relatively less effective. In the same paper the authors provide some empirical evidence, using SJTs that employed scenarios deliberately designed to tap into the constructs of agreeableness, conscientiousness and extraversion. In this sense SJTs may be considered a way of estimating aspects of ITPs, which are not easily directly measurable.

The potential for an SJT to effectively capture an ITP will depend on creating depicted scenarios 'activate' the relevant trait in the respondent. Thus, 'trait-activation theory' (28) has played a key role in informing the development of construct-driven SJTs. This combines *trait theory* (29), '*situationism*' (30) and *personality-job fit* theory (31). Trait theory assumes that individuals have relatively stable 'personalities', emphasising that such predispositions give rise to relatively consistent ways of responding to the world. It is this area of psychological research that is mainly responsible for defining and measuring differing personality traits. A familiar result of such work is the OCEAN model of personality, representing the dimensions of *openness* (to experience), *conscientiousness, extraversion*, *agreeableness* and *neuroticism*. In contrast, 'situationists' argue that it is external situations, rather than relatively stable, underlying traits, that largely influence behavior. These apparently opposing views have largely been reconciled via an interactionist perspective that presumes both traits and situations play a role in

determining the likelihood of an individual responding in a certain way. Their relative contribution would be seen as shifting depending on the particular circumstances observed. Indeed, it is easy to conceptualise how traits may play a stronger role in determining behavior in some circumstances and *visa versa*. *Personality-job fit* theory, as the name suggests, supposes that individuals with combinations of certain traits may be better suited to some work roles than others (32). Thus, a better fit would be assumed to lead to higher job satisfaction and workplace effectiveness. Importantly, recently evidence has been provided that suggests that prosocial personality traits may substantially explain performance differences in SJTs that depict interpersonal scenarios (33). This implies that even traditional SJTs may be inadvertently measuring such traits, albeit less precisely than construct-driven instruments.

Returning to the concept of ITPs, it is clear that SJTs tap into the relevant trait and capture pertinent (imagined) behavioural responses that are the most likely to predict future workplace effectiveness. SJTs that purport to evaluate ITPs, as with traditional selection SJTs, also claim to measure specific aspects of procedural knowledge. The key difference to highlight here is that, unlike traditional SJTs, this procedural knowledge is postulated as being closely related to defined traits or constructs. Of course, there is no guarantee that a candidate who knows what should be done in a particular situation will repeat that behaviour in real life. Indeed, there are a number of factors that may determine the probability that an ITP is manifest in the workplace, including experience and personality traits (27). For example, a candidate may believe that being 'agreeable' is important an individual who is high on the trait of agreeableness is also more likely to actually exhibit this in practice. However, if the person has extensive experience of being in a workplace where agreeableness does not result in desirable outcomes then they may become less likely to exhibit it in the future! Nevertheless, it should be emphasised that ITPs themselves cannot be conceptualised as specific traits, although they are closely related to them. Consequently, SJTs may be considered a way of estimating aspects of ITPs, which are not directly measurable. The idea that such SJTs evaluate ITPs has also been echoed in relation to their use in medical selection (34).

Construct-driven SJT development

Applying this theoretical framework to SJT design, the focus of item development in a construct-driven approach is to create test material that will elicit the trait of interest. That is, such 'latent traits' are assumed to be unobservable until there is an interaction with the external world, which might include a questionnaire item or particular task. At this point the trait becomes manifest, for example, by the candidate selecting a particular response category from a questionnaire item. Of course, the relationship between the level of latent trait possessed by an individual and the observed response is stochastic, rather than deterministic. That is, a certain response will be observed, given the level of the trait being evaluated, with a certain probability that will lie somewhere between zero and one. It is this probabilistic relationship that is modelled using Item Response Theory (IRT), which provides a valuable framework for understanding the measurement model for a trait (see below).

In contrast to traditional selection SJTs, whilst SMEs may contribute to the creation of test material it is primarily psychologists who lead on developing depicted situations and response options for construct-driven SJTs. When selecting scenarios to be used in construct-driven SJTs a 'taxonomy of situations' may be created (35, 36). For example, there may be a series of situations that are likely to involve some degree of inter-personal response that may involve empathy. Categorising potential scenarios may facilitate the process of portraying situations that elicit the desired trait. This process is likely to lead to the inclusion of scenarios, used as the item stems, that are, on average, shorter, and more generic, than those used in traditional selection SJTs (37). That is, they are less detailed and contextualised around specific workplace or educational situations. One of the goals of this development process is to create a set of items that behave in a more 'unidimensional' manner. The dimensionality will inevitably influenced by the scenario, in terms of the traits activated, but also determined by the response options created. In the latter case, responses should be constructed that, ideally, would tap into a single trait. Moreover, Lievens and Guenole stress that, in this context, response options should represent varying degrees of a specific construct, rather than qualitatively different categories (9, 37). For example, looking at the SJT sample included in Figure 4, it can be seen that all the depicted behaviours in the response section are intended to tap into the

trait of ‘dependability’. Note also the use of a Likert scale format is intended to capture differing levels of this construct in the respondent. This approach contrasts with the ‘blueprinting’ process mentioned earlier in traditional SJT development processes, which merely aims to ensure that a sufficient variety of scenarios are included that cover the domain of interest tap into the range of pertinent traits reported by SMEs. Indeed, it is recognised, that in traditional SJTs, even single items may tap into more than one ability in a candidate (38).

When considering the dimensionality of responses to SJTs, we posit that if scenarios are contextualised to a greater degree, a particular item response may be increasingly influenced by other traits, or indeed specific knowledge about a work role, rather than generalised domain knowledge, learned via socialisation (39). Thus, it is probable that SJTs with more contextualised scenarios presented, would be less likely to exhibit unidimensional scoring patterns, where the variance in scores could largely be explained by a single trait. An example of a construct driven SJT item is shown in Figure 4, taken from a test evaluating ‘dependability’, as a facet of conscientiousness (40).

This design process, depicted in Figure 3, has implications for the scoring scheme employed. In traditional SJTs, the scores allocated to response categories for each item are generally determined by their similarity to those derived from the SME panel. In contrast, construct-driven SJT items are generally scored in a similar way to personality assessments; i.e. responses deemed to be associated with higher levels of the trait under examination are allocated higher scores, as in a Likert-type rating scale approach. However, it is also possible to score such SJTs using the traditional SJT approach employing SMEs, or ‘Wisdom of the Crowd’ approaches, based on responses at piloting. The former can be conceptualised as ‘trait’ whilst the latter ‘effectiveness’ scores, with both estimating related, though different, characteristics of the respondent (41). Issues relating to response instructions (‘*would do*’ vs ‘*should do’*) would seem intuitively apply to equally to traditional and construct-driven SJTs. Guenole also offers a more detailed five step procedure for developing construct-driven SJTs (9). According to this process; 1) scenarios are carefully developed by psychologists, then 2) analysed and categorised according to the trait they predominantly tap into, before 3) response alternatives are generated,

then 4) the response instructions are decided on and a scoring key devised before finally 5) creating the test forms and scoring interpretation guidelines for users.

The differing approaches to traditional versus construct-driven test development are likely to lead to some key differences in the way the resultant scores behave, and these will now be discussed.

Traditional vs construct-driven SJTs: psychometric properties

Test 'reliability' and 'information'

The issue of 'internal consistency', or reliability, of SJT items is a cause of frequent confusion in the literature, and requires comment. Traditionally, conventional reliability metrics (e.g. Cronbach's alpha) mainly provide validity evidence for the internal structure of a test (42). Unidimensionality in this sense merely implies to what extent do the items in the test measure the same construct? Therefore any departure from unidimensionality of the test responses will inevitably impact on such metrics of reliability. If traditional selection SJTs tap into numerous traits, as reflected in the original blueprint, then this would be reflected in lower reliability values. Indeed, one meta-analysis of traditional SJTs reported a pooled alpha coefficient value of around 0.46 (43). In contrast, most high stakes tests that measure a single construct would be expected to have values of 0.7 or above. These issues relating to dimensionality and reliability also lead to difficulties with interpreting the meaning of scores derived from traditional selection SJTs (9): what does it say about a candidate's characteristics if they score higher on such a test compared to a low performing test-taker?

The dimensionality of a test is usually assessed by some form of factor analysis. This procedure explores to what extent the observed variations in item scores can be explained by a smaller number of unobserved (latent) traits, or 'factors'. However, there have been well documented difficulties applying such approaches to SJT responses, with such evaluation of traditional SJTs often returning a picture one might describe as 'fuzzy unidimensionality', more formally referred to as 'essential unidimensionality' (44). In this situation there is a general factor that explains a substantial proportion of the variance in many, often most of the test items, but there are also smaller factors which a minority of the items may also load on to.

Dimensionality, and the relative influence of specific scenarios used in the test, can also be explored using extensions of factor analysis, such as the multi-trait multi-method matrix (MTMM) approach (45). This seeks to understand the proportion of variance in responses explained by scenario, rather than trait-level effects. Prior research using this approach reported that the major source of variance seems to be the situations used (46). However, a separate study, using generalisability ('G') theory (see below) reported that a single, dominant, trait accounted for the largest portion of variance in an SJT's scores (47).

Apart from the heterogeneous item content, the low reliability of traditional SJTs may be due in part to the partial scoring models frequently used, with the aim not necessarily being to discriminate amongst candidates for all items. This means that inter-item correlations may be relatively low. By definition, this situation leads to low Cronbach's alpha values. Thus, it can be argued the traditional notion of reliability may not be applicable to such selection SJTs and the use of such statistics is misleading. Consequently alternative approaches to evaluating the reproducibility (consistency) of traditional selection SJT scores have been suggested (48). These include placing more emphasis on the test re-test reliability value, indicating the temporal stability of scores. in contrast to traditional SJTs, early work evaluating the internal consistency of construct-driven SJTs tend to report reliability coefficients that would tend to be deemed acceptable in high stakes tests. For example, an SJT aimed at evaluating 'dependability' reported a Cronbach's alpha value of 0.78. Likewise, an SJT developed to evaluate narrow facets of the personality reported alpha values ranging from 0.55 to 0.75 for the five dimensions (49).

A comprehensive approach to understanding the 'reliability' (reproducibility) of test scores is provided by G theory (50). In G theory several multiple sources of variance in an assessment process are distinguished and the proportion attributable to each can be estimated. This information can then be used to predict the future reproducibility of test scores, using differing assessment designs. Unlike factor analysis and IRT, G theory makes weaker assumptions about the trait or ability under evaluation. G theory proposes that assessments randomly sample items relating to a domain of knowledge or skills defined and selected by the test developer. Thus, especially when dealing with smaller candidate numbers (i.e. under

100) and an assessment that is likely to tap into multiple traits, as might be the case in locally developed test, the latter approach may be more appropriate than the former. In contrast, IRT relies on having larger numbers of test-takers (except in the case of the simplest 'Rasch model') and a clear understanding of the dimensions underlying the responses, so may be less well suited to this context. Thus, in the case of SJTs it would be possible to use a 'G (generalisability) study' to isolate and estimate the sources of variance. This would then be followed up by a 'D' (decision) study that could be used to select the optimum number of scenarios and items, in order to maximize the generalisability (reliability) of the test, for the intended purpose.

When considering how the resultant scores from a test may be best used in the selection process it can also be helpful to consider the concept of 'test information' that has emerged from the IRT tradition. At this point a brief explanation of IRT should be given though a more extensive summary is provided by Reise et al. (51, 52). In IRT the probability that a certain response category will be selected by a test-taker is modelled as a function of the candidate's ability in combination with the characteristics of that particular item. In the simplest form of IRT, the one-parameter logistic, or Rasch model (53), this probability is determined by only two factors; the ability (or trait level) of the candidate and the relative difficulty of the item. More complex IRT models exist that also account for the ability of items to discriminate between test-takers of different ability levels, as well as how easy an item may be to guess correctly, or conversely, answer incorrectly due to carelessness. Each item in a test can therefore be considered as having its own 'item characteristic curve', which displays the relationship between candidate ability (or trait level, usually denoted as theta) and the probability of selecting a particular category (hence obtaining a specific score) on that question. Combined together, these item characteristic curves create a 'test information curve'. This information curve is able to convey the extent to which the test scores are able to discriminate accurately between candidates across the various ability (or trait) levels. For example, tests that are experienced as generally easy or are competency based assessments, discriminate best between candidates at the lower end of ability. Traditional SJTs evaluating non-academic traits tend to be of this latter type because the aim is to provide a range of possible scenario responses from easy to challenging. The added complexity with creating SJT items that are more challenging is that they often not do

not show a high degree of consensus between SMEs when constructing a scoring key. This is because more complex, subtle, situations frequently divide opinion regarding the best and most effective course of action. Such tests that discriminate relatively well between candidates at the lower end of performance are best suited to 'screen out' applicants that failed to reach a relatively low scoring threshold.

However, 'test information' may be more important if using a cut-point score in order to screen-out certain candidates. It is probably less important when being used to rank candidates in order. Nevertheless, where there is less information to discriminate between candidates of a certain ability or trait range the exact order of the rankings will be less certain.

Predictive validity

The overall predictive validity of traditional SJTs for workplace performance is relatively well established in that meta-analytic studies report, in general, low to moderate correlations (around r=0.3) between the scores derived from such assessments and subsequent ratings of workplace performance (16). Likewise, there is accumulating evidence for the acceptability of SJTs, generally, in the context of medical selection and the predictive relationship with other related criteria (34). In this setting, as expected, there is some validity evidence that the scores from SJTs predict later, 'distal', outcomes more strongly than earlier, proximal ones, especially those that relate to interpersonal functioning (54, 55). Indeed, there may be little relationship between performance on such instruments and academic achievement in the early years of medical undergraduate education (56). Evidence for the validity of SJTs evaluating non-academic traits for medical selection purposes has also been sought by linking the test scores to outcomes that may require a degree of interpersonal competence, such as performance in high fidelity simulations of clinical practice in primary care physicians (23), successful completion of the first stage of postgraduate medical training (22, 57) or performance in the first year as GP (58). Cross-sectional evidence that supports the concept of SJTs measuring constructs relevant to interpersonal functioning also exists, in the form of correlations with multiple mini interviews (MMIs) (59, 60). It should be noted that a separate, systematic review of the validity evidence for SJTs in medical selection is being currently undertaken.

At present, research on the predictive validity of construct-driven SJTs is still emerging. However, results so far have been promising, though it should be noted that none of the studies have been conducted in a high-stakes settings in which coaching and faking effects might be at play. There is evidence from general research that construct-driven SJTs have some ability to predict closely related constructs, many of which could be considered relevant to effective medical practice. For example, scores from an SJT evaluating 'personal-initiative' correlated 0.48 with supervisor ratings of the construct (61). A separate study reported variations in SJT responses could be partly explained by their self-rated 'functional flexibility' (the ability to adapt behavioural responses to different situations)(62). Moreover, there are several examples of construct-driven SJTs that show relatively high levels of convergent and divergent validity (49). For example, a recently developed SJT focused on 'dependability', a core facet of conscientiousness, has been evaluated (40). The scores from the instruments showed moderate correlations (r=0.29) with other measures of dependability, such as self-reported biographical data (for example, asking respondents how often they would take more than a day to return a phone call). These relatively consistent properties also mean that construct-driven SJTs, unlike traditional ones, can be used to test theory (9). However, in practice, an SJT that tapped into a single trait may not be perceived as particularly useful for selecting future medical personnel. For this reason, emerging instruments often cover multiple domains, with each representing a distinct dimension and scale. For example, one construct-driven SJT designed to evaluate 'emotional intelligence' is composed of three scales relating to the constructs of: using your own emotions; sensing other's emotion, and; understanding the emotional context of a situation (63). Thus, selectors could use such resulting scale scores to decide on the desirable profile of applicants, with perhaps minimum thresholds on each of the traits measured. Such multi-dimensional instruments could end up relatively lengthy, though, as with personality self-report measures, it is conceivable that, with experience, shorter versions, composed of the best performing items for each scale, could be created.

It should be noted that there are SJT format tests have been developed in relation to single constructs (e.g. 'integrity') for potential use in medical selection (60, 64).

However, these tests were developed using a traditional approach (Figure 3, Table 1), and do not meet the criteria to be considered construct-driven SJTs (37).

'Faking' effects, resulting from social desirability bias are an ever present threat to the validity of non-academically focused selection assessments. This can be investigated by administering tests under two conditions; one where respondents are asked to 'fake good' and a control group (65). Here a note of concern regarding construct-driven SJTs should be sounded, given they have not been evaluated in high-stakes settings. Moreover, if scenarios are less contextualised then 'knowledge' contribution to the SJT score will be reduced, rendering them more vulnerable to such bias.

SJT approaches and widening participation

Medicine is an academically demanding course with extremely stringent academic selection criteria. One reason for the increasingly stringent entrance criteria is the need for selectors to defend rejecting otherwise strong applicants. There has been an unintended consequence of this trend towards an emphasis on traditional measures of intellectual performance: in general, those from more advantaged socioeconomic groups tend to achieve better high school grades, though may not necessarily translate into subsequent superior performance at university (66). Likewise, albeit to a lesser extent, those from disadvantaged groups may have lower scores on cognitively-based selection assessments (67). Thus, this has created additional impetus for measuring other personal qualities, in order to facilitate widening access to medicine. Indeed, some commentators have suggested that the use of SJTs to evaluate nonacademic attributes may have a substantial role to play in widening participation in medical education (68, 69). However, emerging evidence is somewhat mixed in this respect, with findings suggesting that scores from such instruments may be relatively insensitive to group membership from certain underrepresented populations, but not others (68, 69). SJTs in this context, may be culturally-sensitive, and thus their impact on certain groups should be evaluated during piloting (70). Moreover, if rating response formats are used, especially in construct-driven SJTs, then it is known that 'extreme response style' (a tendency to choose extreme points on a scale) may be more common in individuals self-

identifying as being from certain ethnic groups (71). This too could create inter-group differences in scores.

Traditional selection SJTs	Construct-driven SJTs
Development	
Scenario development led by subject matter experts (SMEs)	Scenario development led by psychologists
Scenarios and responses more contextualised in workplace situations	Scenarios and responses more generic and less contextualised
Scenario creation driven by pragmatic considerations and 'blue printing'	Scenario creation informed by trait-activation theory
Responses may tap into a variety of traits or abilities, even in single item	Responses options designed to tap into a single trait only
'Would do' and 'should do' instructions can be used	'Would do' and 'should do' instructions can also be used
Scoring system based on SME opinion consensus ('effectiveness scoring')	Scoring reflects degree of trait, though 'effectiveness scoring' is also an option
Validated via evaluating relationship with ratings of job performance	Validated via relationship between scores and other measures of construct
Properties	
Responses tend to be multi-dimensional (i.e. 'essential unidimensional')	Responses tend to be unidimensional, at least for each component scale
Low reliability on conventional measures	Relatively high reliability values (e.g. alpha ≥0.70)
Usually uninterpretable structure on factor analysis	Usually have interpretable factor structures
May have greater predictive validity for a specific job performance?	Similar overall predictive validity? Probably more generalised predictive validity?
Lower correlations of scores with personality measures (esp. if 'should do' instructions)	Scores tend to correlate more highly with self-report personality measures
May be less prone to faking (esp. 'should do instructions')	May be more prone to faking and coaching
More difficult to equate different test forms	Easier to equate different test forms

Table 1. A summary of the main characteristics relating to the development and likely properties of traditional versus construct-driven situational judgment tests used for evaluating non-academic attributes during a selection process.

Directions for future research

Lievens has recently proposed an agenda setting out the main priority areas for researching the potential of construct-driven SJTs to improve personnel selection (37). This focused on establishing the psychometric properties of construct driven SJT's, in terms of the dimensionality (and hence reliability), as well as their proneness to faking effects. Importantly, it is still to be established whether the benefits of this new approach to SJT construction offers advantages over "bespoke" traditional approaches, in terms of generalisability to other work settings.

In terms of medical selection, there is no high degree of consensus regarding which characteristics are most desirable in a physician. Moreover, the emphasis and nature of these will undoubtedly vary to some extent according to medical specialism, during later stages of training. It therefore may turn out to be the case that construct-level driven SJTs, which evaluate more generic qualities, such as knowledge of interpersonal effectiveness and integrity, agreeableness and conscientiousness, may be more appropriate at selection into early stages of training. This may be especially true where there has been relatively little exposure to the healthcare workplace. Conversely, more traditional SJT approaches could turn out to have high levels of predictive validity for later stages of training, where there has been greater exposure to real life work situations. Such SJT's may benefit from the additional contextualisation that is normally employed by the traditional approach to development, and usually reduced or absent in the construct driven method.

Regarding the widening participation agenda, further research is also required to understand how SJT construction, including scoring method and implementation, may impact on under-represented groups. It is already recognised that such effects are sensitive to both SJT construction and scoring approaches (15, 18).

Conclusion

The use of the SJT format has a long history in personnel selection, though more widespread use in medical education settings is relatively recent. Developers and users of such selection assessments need to be mindful of the rather pragmatic, tradition from which this selection approach has emerged. Whilst many questions

remain over the construct-driven approach to SJT development it may prove useful in the earlier stages of medical selection. Understanding the likely advantages and limitations of both approaches will help test developers and selectors make informed decisions about which approach may be most effective in a given context.

Funding

PAT is supported in his research by an NIHR Career Development Fellowship. This paper presents independent research part-funded by the National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

LWP's research time is part funded by the UCAT Board.

This article was produced as an output from an international workshop on SJTs in medical selection supported by a Worldwide University Network (WUN) Research Development Funding (RDF) award.

Author Contributions

PAT led on conception and drafting of this work. All authors substantially contributed to conception, design, drafting and critically appraising the work. All authors approve the version of the manuscript submitted. All authors agree to be accountable for all aspects of the work.

Competing interest

None

Ethical approval

Not applicable

Figure legends

Figure 1. An example of a situational judgment test item using a ranking response format.

Figure 2. An example of a situational judgment test item using a rating scale response format (used with kind permission from the UCAT Board).

Figure 3. The traditional and construct-driven approaches to developing Situational Judgment Tests for personnel selection**.**

Figure 4. Example of an item from a construct-driven Situational Judgment Test, evaluating 'dependability' (40)

References

1. McDaniel MA, Psotka J, Legree PJ, Yost AP, Weekley JA. Toward an understanding of situational judgment item validity and group differences. J Appl Psychol. 2011;96(2):327-36.

2. Carrington DH. Note on the Cardall Practical Judgment Test. J Appl Psychol. 1949;33(1):29-30.

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

4. Cleland J, Dowell J, McLachlan JC, Nicholson S, Patterson F. Identifying best practice in the selection of medical students. London: General Medical Council; 2012.

5. Tiffin PA, Paton LW, Mwandigha LM, McLachlan JC, Illing J. Predicting fitness to practise events in international medical graduates who registered as UK doctors via the Professional and Linguistic Assessments Board (PLAB) system: a national cohort study. BMC Med. 2017;15(1):66.

6. Association of American Medical Colleges. Situational Judgment Test (SJT) Research 2018 [cited 2019 Sept 20]. Available from: www.aamc.org/admissions/admissionslifecycle/409100/situationaljudgmenttest.html.

7. Arthur W, Jr., Villado AJ. The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. J Appl Psychol. 2008;93(2):435-42.

8. Lievens F, De Soete B. Simulations. In: N.Schmitt, editor. The Oxford Handbook of Personnel Assessment and Selection Oxford: Oxford University Press; 2012. p. 383–410.

9. Guenole N, Chernyshenko OS, Weekly J. On designing construct driven situational judgment tests: Some preliminary recommendations. Int J Test. 2017;17(3):234-52.

10. Bruk-Lee V, Drew EN, Hawkes B. Candidate reactions to simulations and media-rich assessments in personnel selection. In: Fetzer MS, Tuzinski K, editors. Simulations for Personnel Selection New York: Springer; 2014. p. 43-60.

11. Schmitt N, Ostroff C. Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. Pers Psychol. 1986;39(1):91-108.

12. Richman-Hirsch WL, Olson-Buchanan JB, Drasgow F. Examining the Impact of Administration Medium on Examinee Perceptions and Attitudes. J Appl Psychol. 200;85(6): 880-7.

13. Flanagan JC. The critical incident technique. Psychol Bull. 1954;51(4):327-58.

14. Guenole NR, Weekley JA, Ro S, editors. Oral presentation: Scoring Keys and Measurement Models Not Required? SJT Predictions of Job Performance by Recursive Partitioning and its Variations. Oral Presentation. . 31st Annual Conference of the Society for Industrial and Organizational Psychology 2016; Anaheim, CA.

15. De Leng WE, Stegers-Jager KM, Husbands A, Dowell JS, Born MP, Themmen APN. Scoring method of a Situational Judgment Test: influence on internal consistency reliability, adverse impact and correlation with personality? Adv Health Sci Educ. 2017;22(2):243-65.

16. McDaniel MA, Hartman NS, Whetzel DL, Grubb WL. Situational judgment tests, response instructions, and validity: a meta-analysis. Pers Psychol. 2007;60(1):63-91.

17. Nguyen NT, Biderman MD, McDaniel MA. Effects of Response Instructions on Faking a Situational Judgment Test. Int J Select Assess. 2005;13(4):250-60.

18. Lievens F, Sackett PR, Dahlke JA, Oostrom JK, De Soete B. Constructed response formats and their effects on minority-majority differences and validity. J Appl Psychol. 2018.

19. de Leng WE, Stegers-Jager KM, Born MP, Themmen APN. Faking on a situational judgment test in a medical school selection setting: Effect of different scoring methods? Int J Select Assess. 2019;0(0).

20. Christian MS, Edwards BD, Bradley JC. Situational Judgment Tests: Constructs Assessed and a Meta-Analysis of their Criterion-Related Validities. Pers Psychol. 2010;63(1):83-117.

21. Patterson F, Denney M-L, Wakeford R, Good D. Fair and equal assessment in postgraduate training? A future research agenda. Br J Gen Pract. 2011;61(593):712-3.

22. Smith DT, Tiffin PA. Evaluating the validity of the selection measures used for the UK's foundation medical training programme: a national cohort study. BMJ open. 2018;8(7).

23. Patterson F, Tiffin PA, Lopes S, Zibarras L. Unpacking the dark variance of differential attainment on examinations in overseas graduates. Med Educ. 2018;52(7):736-46.

24. Patterson F, Cousans F, Edwards H, Rosselli A, Nicholson S, Wright B. The Predictive Validity of a Text-Based Situational Judgment Test in Undergraduate Medical and Dental School Admissions. Acad Med. 2017.

25. McDaniel MA, List SK, Kepes SJI. The "hot mess" of situational judgment test construct validity and other issues. Ind Organ Psychol. 2016;9(1):47-51.

26. Lievens F, Sackett PR. Situational Judgment Tests in High-Stakes Settings: Issues and Strategies With Generating Alternate Forms. J Appl Psychol. 2007;82(4):1043-55.

27. Motowidlo SJ, Hooper AC, Jackson HL. Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. J Appl Psychol. 2006;91(4):749-61.

28. Tett RP, Burnett DD. A personality trait-based interactionist model of job performance. J Appl Psychol. 2003;88(3):500-17.

29. Allport GW. Concepts of trait and personality. Psychol Bull. 1927;24(5):284-93.

30. Upton CL. Virtue Ethics and Moral Psychology: The Situationism Debate. J Ethics. 2009;13(2-3):103-15.

31. Chatman JA. Improving interactional organizational research: A model of person-organization fit. Academy Manage Rev. 1989;14(3):333-49.

32. Judge TA, Zapata CP. The Person–Situation Debate Revisited: Effect of Situation Strength and Trait Activation on the Validity of the Big Five Personality Traits in Predicting Job Performance. Academy Manage J. 2015;58(4):1149-79.

33. Motowidlo SJ, Lievens F, Ghosh K. Prosocial implicit trait policies underlie performance on different situational judgment tests with interpersonal content. Hum Perform. 2018;31(4):238-54.

34. Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. Med Teach. 2016;38(1):3-17.

35. Parrigon S, Woo SE, Tay L, Wang T. CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. J Pers Soc Psychol. 2017;112(4):642-81.

36. Rauthmann JF, Gallardo-Pujol D, Guillaume EM, Todd E, Nave CS, Sherman RA, et al. The Situational Eight DIAMONDS: a taxonomy of major dimensions of situation characteristics. J Pers Soc Psychol. 2014;107(4):677-718.

37. Lievens F. Construct-Driven SJTs: Toward an Agenda for Future Research. Int J Test. 2017;17(3):269-76.

38. Sorrel MA, Olea J, Abad FJ, de la Torre J, Aguado D, Lievens F. Validity and Reliability of Situational Judgement Test Scores:A New Approach Based on Cognitive Diagnosis Models. Organ Res Methods. 2016;19(3):506-32.

39. Lievens F, Motowidlo SJ. Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. Ind Organ Psychol. 2016;9(1):3-22.

40. Olaru G, Burrus J, MacCann C, Zaromb FM, Wilhelm O, Roberts RD. Situational Judgment Tests as a method for measuring personality: Development and validity evidence for a test of Dependability. PLoS One. 2019;14(2):e0211884.

41. Motowidlo SJ, Hooper AC, Jackson HL. Situational judgment tests: Theory, measurement, and application (pp. 57–82). Erlbaum, NJ: Lawrence Erlbaum. A theoretical basis for situational judgment tests. In: Weekley JA, Ployhart RE, editors. Situational judgment tests: Theory, measurement, and application. Mahwah, NJ: Lawrence Erlbaum; 2006. p. 57–82.

42. Sijtsma K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. Psychometrika. 2009;74(1):107-20.

43. Catano VM, Brochu A, Lamerson CD. Assessing the reliability of situational judgment tests used in high-stakes situations. Int J Select Assess. 2012;20(3):333-46.

44. Nandakumar R. Traditional Dimensionality Versus Essential Dimensionality. J Educ Measure. 1991;28(2):99-117.

45. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959;56(2):81-105.

46. Westring AJF, Oswald FL, Schmitt N, Drzakowski S, Imus A, Kim B, et al. Estimating Trait and Situational Variance in a Situational Judgment Test. Hum Perform. 2009;22(1):44-63.

47. Jackson D, Lopilato A, Hughes D, Guenole NR, Shalfrooshan A. The Internal Structure of Situational Judgment Tests Reflects Candidate Main Effects: Not Dimensions or Situations. J Occup Organ Psychol. 2016.

48. Catano V, Brochu A, D. Lamerson C. Assessing the Reliability of Situational Judgment Tests Used in High-Stakes Situations. Int J Select Assess. 2012;20(3):333-46.

49. Mussel P, Gatzka T, Hewig J. Situational Judgment Tests as an Alternative Measure for Personality Assessment. Eur J Psychol Asses. 2016:1-8.

50. Cronbach LJ, Rajaratnam N, Gleser GC. Theory of Generalizability: A Liberalization of Reliability Theory. Br J Stat Psychol. 1963;16(2):137-63.

51. Abdel Wahed WY, Hassan SK. Prevalence and associated factors of stress, anxiety and depression among medical Fayoum University students. Alex J Med. 2016. http://dx.doi.org/10.1016/j.ajme.2016.01.005.

52. Reise SP, Ainsworth AT, Haviland MG. Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research. Curr Dir Psychol Sci. 2005;14(2):95-101.

53. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press; 1960.

54. Lievens F, Buyse T, Sackett PR. 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 Pyschol. 2005;90(3):442-52.

55. Paton LW, Tiffin PA, Smith D, Dowell JS, Mwandigha LM. Predictors of fitness to practise declarations in UK medical undergraduates. BMC Med Educ. 2018;18(1):68.

56. Tiffin PA, Paton LW. Exploring the validity of the 2013 UKCAT SJT- prediction of undergraduate performance in the first year of medical school: Summary Version of Report. The UKCAT Board; 2017.

57. Roberts C, Khanna P, Rigby L, Bartle E, Llewellyn A, Gustavs J, et al. Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45. Med Teach. 2018;40(1):3-19.

58. Lievens F, Sackett PR. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. J Appl Psychol. 2012;97(2):460-8.

59. Patterson F, Rowett E, Hale R, Grant M, Roberts C, Cousans F, et al. The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. BMC Med Educ. 2016;16:87.

60. Husbands A, Rodgerson MJ, Dowell J, Patterson F. Evaluating the validity of an integrity-based situational judgement test for medical school admissions. BMC Med Educ. 2015;15:144.

61. Bledow R, Frese M. A situational judgment test of personal initiative and its relationship to performance. Pers Psychol. 2009;62(2):229-58.

62. Lievens F, Lang JWB, De Fruyt F, Corstjens J, Van de Vijver M, Bledow R. The predictive power of people's intraindividual variability across situations: Implementing whole trait theory in assessment. J Appl Psychol. 2018;103(7):753-71.

63. Sharma S, Gangopadhyay M, Austin E, Mandal MK. Development and validation of a situational judgment test of emotional intelligence. Int J Select Assess. 2013;21(1):57-73.

64. Leng WE, Stegers-Jager KM, Born MP, Themmen APN. Integrity situational judgement test for medical school selection: judging ‘what to do’ versus ‘what not to do’. Med Educ. 2018;52(4):427-37.

65. McFarland LA, Ryan AM. Variance in faking across noncognitive measures. J Appl Psychol. 2000;85(5):812-21.

66. Mwandigha LM, Tiffin PA, Paton LW, Kasim AS, Böhnke JR. What is the effect of secondary (high) schooling on subsequent medical school performance? A national, UK-based, cohort study. BMJ open. 2018;8(5).

67. Tiffin PA, McLachlan JC, Webster LAD, Nicholson S. Comparison of the sensitivity of the UKCAT and A Levels to sociodemographic characteristics: a national study. BMC Med Educ. 2014;14:7.

68. Lievens F, Patterson F, Corstjens J, Martin S, Nicholson S. Widening access in selection using situational judgement tests: evidence from the UKCAT. Med Educ. 2016;50(6):624-36.

69. Juster FR, Baum RC, Zou C, Risucci D, Ly A, Reiter H, et al. Addressing the Diversity-Validity Dilemma Using Situational Judgment Tests. Acad Med. 2019.

70. Lievens F, Corstjens J, Sorrel M, Abad F, Díaz J, Ponsoda V. The Cross-cultural Transportability of Situational Judgment Tests: How does a US-based integrity situational judgment test fare in Spain? Int J Select Assess. 2015:23(4). 361-72..

71. Peterson RA, Rhi-Perez P, Albaum G. A Cross-National Comparison of Extreme Response Style Measures. Int J Mark Res. 2014;56(1):89-110.