Australasian Journal of Organisational Psychology, Volume 8, e7, 1–10
© Australasian Psychological Society Ltd 2015. doi 10.1017/orp.2015.6

Best Practice Recommendations for Situational Judgment Tests

Stefanie Pollard[1] and Helena D. Cooper-Thomas[2]
[1] *Talegent, Auckland, New Zealand*
[2] *School of Psychology, The University of Auckland, New Zealand*

Situational judgement tests (SJTs) are increasingly popular as part of selection procedures because of their strong psychometric properties. SJTs present test takers with job-related situations and a series of potential responses. The aim of this review article is to provide evidence-based recommendations for practitioners who are considering implementing SJTs for their clients, as well as offering an analysis of some key issues for academics. This research provides a review of the literature on SJTs to answer three questions: (1) What is the difference between framing questions either as what *should* you do versus what *would* you do? (2) What is the difference between multimedia-based and text-based SJTs? (3) What is the best process for SJT creation? Following this, recommendations are provided as follows: That responses are optimally framed in terms of what a respondent thinks s/he *should* do, not what s/he *would* do; that multimedia-based SJTs are used in preference to text-based SJTs; and a 'best practice' process for creating an SJT is outlined. Finally, the review includes a discussion of implications for creating tests that are inclusive of different ethnic groups, as well as ethical issues relating to the use of SJTs in selection.

■ **Keywords** situational judgment test, SJT, selection, item stem, validity

Situational judgment tests (SJTs) are a form of psychometric assessment whereby a candidate is presented with a job-related situation describing a problem, with a range of possible responses to select from (Lievens, Peeters, & Schollaert, 2008). Typically, SJTs are used to simulate the kinds of scenarios that employees would experience at work and are therefore part of a larger suite of work simulation assessments (McDaniel & Nguyen, 2001). Work simulations, in general, show high face validity relative to other types of assessment, due to their simulation fidelity. That is, they are viewed by candidates as similar to the target job and work situation (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993). Moreover, a meta-analysis across 17 countries found they were rated by candidates as one of the fairest assessment methods (Anderson, Salgado, & Hulshegger, 2010).

Several meta-analyses of SJTs have been conducted and provide criterion validity data. McDaniel, Morgeson, Finnegan, and Campion (2001) found corrected criterion validities of SJTs with performance of .18 for predictive designs and .35 for concurrent designs. McDaniel, Hartman, Whetzel, and Grubb (2007) found the same criterion validity of .26 across two types of SJTs — knowledge and behavioural tendencies. More recently, Christian, Edwards, and Bradley (2010) showed that corrected criterion validities with job performance ranged from .19 to .43, depending on which facets of SJTs were examined. In their study, SJT dimensions relating to job knowledge and skills showed the lowest criterion validity for job performance ($r = .19$), whereas personality composites ($r = .43$) and teamwork skills ($r = .38$) had the highest criterion validities with performance.

Part of the benefit of SJTs over other forms of assessment is their flexibility. While a large amount of empirical research indicates that SJTs are valid indicators of job performance (e.g., Lievens, Sackett, & Buyse, 2009), they may also be used as recruitment tools that yield more positive candidate perceptions of the organisation relative to other types of assessment (Bauer & Truxillo, 2006), especially when provided in multimedia format

ADDRESS FOR CORRESPONDENCE: Helena D. Cooper-Thomas, School of Psychology, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. Email: h.cooper-thomas@auckland.ac.nz

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

(Chan & Schmitt, 1997; Olson-Buchanan & Drasgow, 2006). Moreover, where organisations use competency models that overlap across roles and levels, a single SJT can be used for a wide variety of roles. This offsets the potentially higher development costs through greater applicability relative to other assessment methods.

The evidence suggests that SJTs hold great potential as an assessment method, but for practitioners wishing to use SJTs, there are few resources to guide them. Therefore, the aim of this critical review is to provide evidence-based recommendations for practitioners who are considering implementing SJTs for their clients, as well as offering a review of some key issues for academics. This aligns with earlier calls for pragmatic science as the optimal way for our field to proceed (Anderson, Herriot, & Hodgkinson, 2001), and more recent calls for evidence-based research that incorporates practitioner voices (Briner & Rousseau, 2011). In line with these appeals for better translation of research to practice, our focus was to answer a set of practice-relevant questions through a review of past SJT research, written by a mixed practitioner-academic team.

With regard to SJT research, we note the relative dearth of peer-reviewed published work on SJTs that directly helps practitioners to implement them, in spite of their being one of the most valid forms of assessment. Various factors may contribute to this lack of quality research (Briner & Rousseau, 2011), including practitioners implementing SJTs but not writing up their research due to the lack of incentives for publication; or academics being reluctant to publish in such a constrained area of practice. In particular, since lack of theory can hinder publication (Grant & Pollock, 2011), combined with the fact that SJT research has been more practical than theoretical (Campion, Ployhart, & MacKenzie, 2014; Chan, 2006), this may deter academics from SJT research. With many practitioners previously being students and relying on taught courses for their organisational psychology knowledge, a lack of review material that concisely presents SJT findings could hinder academics from teaching SJT material. Hence, we hope also that the present critical review will facilitate both the development of the knowledge base and also inform practice both for current practitioners and for academics teaching future practitioners.

There have been two reviews of SJTs in recent years (Campion et al., 2014; Christian et al., 2010), and it is important to identify the distinct contribution of the present research. Christian et al. (2010) conducted a meta-analysis focused on the constructs measured by SJTs, and identified the criterion-related validities of these. This is practically useful since practitioners can focus their SJT development efforts on those constructs with the highest criterion-related validities. However, this information does not inform practitioners on issues of SJT development. The more recent quantitative, descriptive summary provided by Campion et al. (2014) identifies the most common formats and development methods for SJTs, such as response format, but does not evaluate their benefits or drawbacks in practice. Thus, our review differs by having a narrow focus on three questions that practitioners wanting to leverage SJT research will want answers to. Our three questions used to interrogate past research are: (1) What is the difference between framing questions as what *should* you do versus what *would* you do? (2) What is the difference between multimedia-based and text-based SJTs? (3) What is the best-practice process for creating an SJT?

Method

This section outlines the search strategy and selection criteria used for this review, including descriptions of the types of studies reviewed. The review was restricted to published academic peer-reviewed articles within industrial and organisational psychology, management, and human resources journals, and book chapters that are referenced in such articles.

First, articles were sourced using a Google Scholar search.[1] There were three key research questions to be answered, and we selected separate sets of search terms for each. Given that SJT research is fairly sparse, we imposed no date range limitations.

To answer the question 'What is the difference between framing questions as what should you do versus what would you do?', the following search terms were used: 'situational judgment test' and 'instructions', 'situational judgment test' and 'should', 'situational judgment test' and 'response'. We looked for articles that appeared to have the potential to answer the question above. If a review of the abstract indicated that the article was unlikely to address the question above, it was not included. This resulted in five articles. A further three articles and/or book chapters were identified from a review of the reference lists of these articles.

At this point, we assessed whether the resources assembled so far were sufficient to answer the second question: 'What is the difference between multimedia-based and text-based SJTs?' On determining that our resources were insufficient, we continued the search using the following search terms: 'situational judgment test' and 'video', 'situational judgement test' and 'multimedia'. As before, if the abstract indicated that the article was unlikely to address the question above, it was not included. We identified three additional articles through this direct search, and two further articles and book chapters by scanning the reference lists of these.

Our final question was 'What is the recommended process for creating a situational judgment test?' We

[1] We repeated our search using PsycInfo with the same combinations of "situational judgment" and "should", and "instructions", and "response" respectively. This provided a far smaller number of resources — of 3, 0 and 5 articles for these combinations respectively compared with Google Scholar which stated "about" 616, 1,090 and 1,030 results respectively. Therefore we used Google Scholar for our searches as providing a more comprehensive search function.

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

determined that we had gathered sufficient information in the articles and chapters so far to answer this question.

While writing this review for publication, two additional articles were published and subsequently incorporated (Campion et al., 2014; St-Sauveur, Girouard, & Goyette, 2014). All book chapters and journal articles used for the literature review have been marked with an asterisk in the references section of this article. This represents a total of 13 articles and/or book chapters. While this is a small number, Campion et al. (2014) found only 59 articles relevant to all aspects of SJTs when searching from 1990 onwards. A closer comparison can be drawn with the research of Arthur et al. (2014). Their SJT question focused on the response formats of SJT items as an important design feature, and their 'thorough search of the extant literature' (p. 536) yielded a single study on response formats (Ployhart & Erhart, 2003). Hence, our identification of 13 articles relevant to the three SJT questions posed is not unusual, and reflects the lack of research.

Results

Impacts of Framing Questions Either as What Should You Do or What Would You Do?

One key variation in SJTs is whether they ask test takers to select what they think is the best answer ('What should you do?') or whether they ask test takers to select what they personally would do in the outlined situation ('What would you do?'; Lievens et al., 2008). SJTs that ask test takers to indicate what they should do can be considered a measure of knowledge as test takers are being asked to demonstrate that they know what the most effective option is (McDaniel et al., 2007). Thus, they are a measure of maximal performance (McDaniel et al., 2007); in other words, the best possible performance for that individual. Lievens et al. (2008) suggested that this makes 'should' item formats congruous with cognitive items, and henceforth we describe this format as knowledge-based items. On the other hand, when the item asks test takers what they would do, it is assessing behavioural tendency and test takers are expected to report how they tend to behave, sharing similarities with personality questionnaires (Lievens et al., 2008). We describe this format as behavioural-based items.

Research supports the existence of framing effects, with different instruction formats resulting in different correlations between cognitive assessments. For example, in the meta-analysis conducted by McDaniel and colleagues (2007), SJTs with knowledge-based instructions were found to correlate more strongly with cognitive ability tests (.35) than behavioural-based SJT items (.19). Furthermore, behavioural-based SJT items correlated more strongly with personality measures, achieving correlations of .37 with agreeableness, .35 with emotional stability, and .34 with conscientiousness, in comparison with knowledge-based instructions, which achieved correlations of .19, .12, and .24 respectively.

It seems likely that item formats will affect the amount of response distortion by SJT test takers, particularly in high-stakes testing environments such as when applying for a job (Lievens et al., 2009). Behavioural-based SJTs are likely to be more affected by impression management and self-deception than are knowledge-based items. For example, when answering a behavioural-based item, those with tendencies to deceive themselves may select an answer that demonstrates more favourable behaviour than what they may actually do in reality. Alternatively, a test taker may recognise the behavioural-based item that they are most likely to perform but choose to select an alternative response that they believe would get them a higher score on the SJT, that is, choose to 'fake good' (McDaniel et al., 2007). In contrast, knowledge-based items may be more difficult to fake as this requires sufficient knowledge to recognise the best answer.

There is some empirical research supporting the idea that behavioural-based SJTs are more susceptible to faking. Using a sample of students, Nguyen, Biderman, and McDaniel (2005) asked one group to answer an SJT honestly and another group to fake good. Each group completed both a behavioural-based SJT and a knowledge-based SJT. Participants responding to behavioural-based SJT were able to fake good (effect sizes of .15 to .34) whereas the knowledge-based items were viewed as relatively impervious to faking. However, this study did not examine whether test takers in high-stakes situations do respond differently to the two instruction formats.

Subsequent research has investigated a high-stakes situation, namely entry into a university course that was based on SJT scores. Using a large group of medical school applicants, Lievens and colleagues (2009) randomly assigned participants to complete either a behavioural-based or knowledge-based SJT. They found no significant differences between the scores of these two groups. Consistent with McDaniel and colleagues (2007), they also found that the knowledge-based SJT had a higher correlation with cognitive ability scores than did the behavioural-based SJT (.19 in comparison to .11). This was smaller than the difference found by McDaniel and colleagues and suggests that the format of response instructions does not have a large impact on response distortion in practice. Lievens and colleagues (2009) still concluded that knowledge-based instructions may be superior as candidates do not have to consider whether or not they should answer honestly, making faking a non-issue.

It is also important to consider the process required for candidates to answer knowledge-based and behavioural-based items. SJTs with both response formats require test takers to make predictions (Brooks & Highhouse, 2006). Knowledge-based items require test takers to predict what

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

would happen if they selected each of the available options and then make a decision as to which option leads to the best outcome. Behavioural-based items require test takers to do more, as they also have to think about their likely behaviour in that situation (Brooks & Highhouse, 2006). This is potentially more challenging since research indicates that people are not very good at predicting their own behaviour. People tend to be overly optimistic when thinking about their future behaviour, and are poor at making predictions about the time needed to complete tasks (Brooks & Highhouse, 2006). Furthermore, people find it difficult to accurately predict how they will feel in the future (Loewenstein & Schkade, 1999), particularly when they are currently in a relaxed state and required to predict how they might feel when they are more agitated (Loewenstein, 1996). This led Brooks and Highhouse (2006) to suggest that SJT developers should consider avoiding questions that require candidates to either make predictions relating to time estimates or to indicate how they might behave when they are in a highly affect-laden situation.

A final crucial issue is whether the different response options for SJTs have differential criterion-related validity. The ability of SJTs to predict job performance is often explained as being due to candidates indicating how they would behave in a job-relevant situation (Brooks & Highhouse, 2006). However, it may be argued that this is only true of behavioural-based SJTs where the candidate indicates their likely actual behaviour.

Several researchers have looked at whether the question format affects criterion-related validity. Ployhart and Ehrhart (2003) compared these formats in a large sample of students and found significant differences. Their research showed that only behavioural-based SJTs were related to the performance criteria assessed, which were grade point average (GPA), SAT total (U.S. college admission score), a self-report measure of performance, and a peer-rated measure of performance. Nonetheless, there was no measure of job performance included in this study, making it difficult to draw firm conclusions on which type of SJT yields better criterion-related validity for work settings (Ployhart & Ehrhart, 2003).

The issue of whether it makes a difference to focus the SJT on knowledge or typical behaviour was further investigated in McDaniel et al.'s (2007) meta-analysis. Intriguingly, they suggested that knowledge-based SJTs assess maximal performance, whereas behavioural-based SJTs assess typical performance. In other words, the construct validity of the test is affected by the response instructions. However, the criterion-related validity in predicting performance is not affected by response instructions, with correlations of .26 both for knowledge-based and behavioural-based SJTs (see also Lievens et al., 2009). Weighting the meta-analytic evidence more heavily, this suggests that the response format does not make a difference to criterion-validity.

Impacts of Text-Based versus Multimedia-Based SJTs

In their original conception, SJTs were text-based. That is, the candidate would be provided with one or more written scenarios and they would select from up to five written response options (Lievens et al., 2008). With the increasing use of technology at work, and concomitantly in selection procedures, SJTs can now be presented in a variety of multimedia formats. These multimedia assessments generally consist of a number of work-related scenarios being presented, followed by the presentation of one or more questions where the test taker is required to select an option from the typically four or five presented choices (Weekley & Jones, 1997). There are two main types of multimedia-based assessments: video assessments and animated assessments. A video assessment typically involves filming either actors or job incumbents, while an animated assessment shows animated characters.

Research suggests that multimedia-based assessments may have some advantages over text-based SJTs. Multimedia-based SJTs are likely to have high fidelity because more information is available to the candidate regarding the scenario, and the situation can be displayed in such a way that it is closer to the performance required on the job (Lievens et al., 2008). This may in turn lead to a stronger correlation with job performance. Watching a multimedia-based SJT can also serve as a realistic job preview for job candidates (Weekley & Jones, 1997). In contrast, a text-based assessment has lower fidelity as less detail and realism can be presented (Lievens et al., 2008).

Comparing test results on a text-based assessment versus a video-based one, Chan and Schmitt (1997) found that test takers scored significantly higher on the video-based assessment. They postulate that this may be because the video-based simulation is more realistic and concrete, allowing test takers to better imagine the situation and thus to make better contextual decisions. Their study suggests that improving the fidelity of the SJT may lead to increased test performance in candidates (Chan & Schmitt, 1997).

Multimedia-based SJTs have also been found to have higher face validity than text-based SJTs. Richman-Hirsch, Olson-Buchanan, and Drasgow (2000) received positive reactions to a multimedia-based SJT on conflict resolution skills, with the multimedia-based SJT being rated as more face valid than either text-based computer-administered assessments or a paper-and-pencil version. Similarly, student participants in Chan and Schmitt's (1997) study rated a video-based SJT as slightly more face valid than a text-based SJT. Importantly, the multimedia-based SJT was also rated to be more enjoyable than the other tests (Richman-Hirsch et al., 2000), suggesting that this format may have benefits for candidate recruitment, as well as motivation for test taking.

Another potential benefit of multimedia-based SJTs is that this format of assessment can assess test takers' interpersonal abilities, such as noticing non-verbal cues and

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

responding appropriately. Through the use of animation or video, multimedia-based SJTs can present scenes to give test takers the ability to see and hear people interacting, allowing them to perceive their emotions and body language and to make decisions on the basis of this. For example, Olson-Buchanan and Drasgow (2006) describe the following situation:

> *a video presentation of an employee complaining about the difficulty of a work assignment might require the assessee to notice and interpret how the employee presents her complaint (e.g., is the employee exasperated or calm?) and this interpretation may be an important factor in selecting the 'right' answer. (pp. 254–255)*

The rich portrayal of such a scene has the potential to add a measure of interpersonal skills to the SJT. That is, emotional cues can be provided subtly and realistically, with interpersonally skilled respondents identifying these. In contrast, it is difficult to portray emotions subtly in a text-based SJT; instead, it would require directly stating the non-verbal cue and thus it would not differentiate between those who did versus those who did not notice these (Olson-Buchanan & Drasgow, 2006).

In line with this sensitivity of SJTs to interpersonal abilities, Lievens and Sackett (2006) found video-based SJTs to have significantly higher predictive and incremental validity for predicting interpersonally oriented criteria than text-based SJTs. Their results suggested that the higher validity of the video assessment was because of the fidelity and closer resemblance of the video assessment to the actual criterion. They also found higher correlations between text-based SJTs and cognitive ability, and higher correlations between the video-based SJT and an interpersonal criterion. They concluded that video-based SJTs are likely to be superior to text-based SJTs when assessing interpersonal skills (Lievens & Sackett, 2006). Christian et al.'s (2010) meta-analysis supports these results, although with the proviso that only small numbers of video-based SJTs were available for analysis. In the case of interpersonal skills, video-based SJTs had corrected criterion-validity of .47 with job performance compared with .27 for paper-and-pencil; for leadership, the corrected criterion-validities were .33 and .27 respectively but with overlapping confidence intervals; for heterogeneous composites (i.e., where a mixed set of knowledge and skills were assessed by the SJT), the corrected criterion-validities were .25 and .36 respectively.

As noted above, text-based assessments provide less rich formats. This can leave them open to more ambiguity as there tends to be less item elaboration than in multimedia-based assessments (Brooks & Highhouse, 2006). There has been very little research relating to SJTs and item elaboration, but when ambiguous items are used, there could be large differences in the ways in which different test takers interpret the items and in the ways in which test creators meant them. For example, it is possible that the test creator has a particular context in mind and identifies the best answer on the basis of this. The test taker may view the ambiguous item in a different way and may well select the best response for their alternative construal of the situation, but which would not be scored as correct (Brooks & Highhouse, 2006). Since a multimedia-based SJT can provide more detail, this may provide greater clarification of the context and hence may be a way of reducing ambiguity and associated error.

Other research indicates that multimedia-based SJTs actually measure slightly different abilities from text-based SJTs (Weekley & Jones, 1997). Text-based assessments have a higher reading requirement as test takers are required to read and interpret the written information before making any decisions. In contrast, video-based assessments require less cognitive ability as they have a lower reading component. This is consistent with the results reported above by Lievens and Sackett (2006) where they found text-based SJTs to correlate more strongly with cognitive ability, and video-based SJTs to correlate more strongly with an interpersonal criterion.

One implication of this is that multimedia-based SJTs have the potential to reduce the adverse impact of testing, particularly for jobs that do not require reading skills (Weekley & Jones, 1997). In support of this, video-based assessments were found to have a smaller difference in testing scores between Black and White test takers when compared with results on a text-based SJT (Chan & Schmitt, 1997). Thus, Chan and Schmitt (1997) found that White test takers scored more highly than Black test takers on both forms of assessment. However, this performance gap was substantially smaller for the video-based SJT. Chan and Schmitt also found the method of assessment to interact with reading comprehension ability, with a moderate correlation found between reading comprehension and performance on the text-based SJT but no significant correlation between reading comprehension and performance on the video-based SJT. Chan and Schmitt suggested that this provides support for the idea that the interaction between reading comprehension and method of testing could partially explain the lower levels of adverse impact for Black test takers on video-based assessments.

Creating a Situational Judgement Test

Developing item content. In creating an SJT, the first step is to develop the item content. Job analysis methods are frequently used, with critical incident interviews as the most common approach to develop SJT content (Motowidlo, Crook, Kell, & Naemi, 2009; Weekley, Ployhart, & Holtz, 2006). Indeed, Campion et al. (2014) note that 82% of studies use critical incidents to design the situation stimuli. SJTs created using job analysis have been found to have significantly higher validity than those that do not (0.38 vs. 0.29; Lievens et al., 2008). The critical incident interview method involves asking subject matter experts (SMEs), typically incumbents and supervisors

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

of incumbents, to recall examples of exceptionally good and exceptionally poor performance. A recommended approach here is the antecedent-behaviour-consequence (or A-B-C) format, whereby the antecedent context is explained (A), then the behavioural actions taken by the employee (B), and then the consequences of the actions are discussed (C). The context or antecedent can be used to create the item stem for the SJT and the employee's behaviour can become one of the response options for the stem (Weekley et al., 2006). Critical incidents can also be gathered from archival records, such as records of accidents (Weekley et al., 2006).

SJTs may also be written to fit a conceptual model (Weekley et al., 2006). This is possible when thorough job analysis information is available that outlines the behavioural dimensions important to job performance. Weekley and colleagues recommend a combination of both — first developing a model either from job analysis or theory, and then consulting with SMEs to add more detail and to identify job relevant situations.

Creating response options. Next, the content for the response options must be created. While there is a paucity of research into the difference between having test developers create item responses and having SMEs create these, Weekley and colleagues (2006) suggest that using SMEs for this stage of development may yield several benefits. First, SMEs will have a good understanding of the kinds of situations faced in the role and thus can create realistic items and response options. Second, consulting with multiple groups of SMEs can result in a pool of various realistic responses to a given situation.

The creation of response options can be done with different groups of SMEs (Lievens et al., 2008). It can also be useful to consult inexperienced employees who can provide responses with differing levels of effectiveness. The test developer can then edit the response options to ensure that they are credible and span a range of effectiveness (Lievens et al., 2008).

Developing a scoring key. The final stage is developing a scoring key, for which there are two differing options: the rational approach and the empirical approach (Lievens et al., 2008). The rational approach involves having current high performing employees rate the effectiveness of the various response options, with the best choices being scored as correct and the worst as incorrect. The empirical approach involves administering the items to a large pilot group and the weightings for each response being determined on the basis of their correlations with a criterion variable such as job performance (Lievens et al., 2008). A combination of these two approaches can also be used (Lievens et al., 2008). There are various means of scoring response options once developed, but these are beyond the scope of the current analysis (for recent work on scoring see St-Sauveur et al., 2014).

Discussion

Framing Questions as Should or Would (Knowledge- Versus Behavioural-Based Items)

Our first question was: 'What is the difference between framing questions as what should you do and what would you do?' Our review indicates that important differences exist between asking test takers what they think they *should* do versus what they *would* do. Asking test takers to indicate what someone should do requires them to think about the best possible option from the alternatives and hence can be considered to be a measure of maximal performance. This makes this format of question impossible to fake good, as a test taker is not being asked to report on his or her own performance but instead to demonstrate job knowledge.

On the other hand, when test takers are asked to indicate what they would do, the question is closer to measures of personality and is at risk of distortion. Specifically, test takers may either fake good or self-deceive, whereby test takers may deliberately or accidentally inaccurately predict how they would act in the situation and select a more effective response than they would display in reality. People are typically quite poor at predicting their behaviour in the future, especially when people in a calm state are asked to indicate how they might act in a stressful or demanding situation (Loewenstein, 1996). In spite of this risk, there is a lack of evidence that test takers do actually distort their answers more when asked to indicate how they would act. However, by avoiding this style of question, there is no risk of faking good. The evidence to date suggests that the two response formats work similarly well in predicting job performance. That is, their criterion validity is similar (Lievens et al., 2009; McDaniel et al., 2007; although see Ployhart & Ehrhart, 2003), even if their construct validity may differ (McDaniel et al., 2007).

Recommendation for framing questions: Use a knowledge-based stem, 'What should you do?'. Overall, the evidence supports asking candidates what they *should* do for several key reasons. First, this method is unaffected by faking good or by forms of self-perception bias. Second, it provides a measure of maximal performance, equal to asking candidates what they would do in terms of predicting job performance. Thus, we recommend that candidates are asked to indicate what the most effective response is to a given situation (what *should* they do).

Differences Between Multimedia-Based Versus Text-Based SJTs

The evidence supports multimedia-based SJTs as superior to text-based SJTs in a variety of areas. Yet research to date suggests that written formats predominate, being found in the majority of published studies (Campion et al., 2014). Multimedia-based SJTs are able to convey

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

information more quickly, meaning they are able to provide an accurate representation of the job situation and can make the situation seem more concrete (Christian et al., 2010). Multimedia-based SJTs allow test creators to produce simulations that are similar to those the candidate will face if hired, acting as a realistic job preview (Weekley & Jones, 1997). Test takers also seem to have a preference for multimedia-based SJTs over text-based SJTs: Multimedia SJTs are rated as being more face valid and more enjoyable to sit (Chan & Schmitt, 1997; Richman-Hirsch et al., 2000), which may have recruitment benefits for organisations.

The ability to show people interacting allows multimedia-based SJTs to include a measure of interpersonal skill, which is more difficult to achieve with a text-based SJT (Christian et al., 2010). Since video-based SJTs have been found to be better predictors of interpersonally oriented criteria than text-based SJTs, they have the additional potential to contribute to incremental validity (Olson-Buchanan & Drasgow, 2006). Text-based SJTs also leave more room for ambiguity (and hence error), which can be better controlled for in a multimedia-based SJT as the test creator is able to choose exactly what is shown. Indeed, this may partly explain why test-takers typically perform better on a multimedia SJT (Chan & Schmitt, 1997).

A final benefit of multimedia-based SJTs is their weaker correlations with cognitive ability. This suggests they have the potential to reduce the adverse impact of testing, as there tend to be smaller differences between cultural groups on multimedia-based SJTs than on text-based SJTs, possibly due to the lowered reading requirements of a multimedia-based SJT.

Contrasting these benefits, one drawback of multimedia-based SJTs should be mentioned, as identified in Campion et al.'s (2014) discussion of SJT reliability. Specifically, video-based SJTs, perhaps exactly because of their richer format, have greater item-specific variance. In turn, this reduces internal reliability. However, as Campion and colleagues and previous authors (e.g., Motowidlo et al., 2009) have emphasised, internal reliability is inappropriate since this provides a measure of unidimensional consistency whereas SJTs are designed to be multidimensional.

Recommendation for format: Use multimedia-based SJTs. Multimedia-based SJTs clearly outperform text-based SJTs in a variety of areas, including ability to predict interpersonally oriented criteria, being less ambiguous, having higher fidelity, higher face validity, and less adverse impact. Thus, we recommend the use of multimedia-based SJTs over text-based SJTs.

How to Develop an SJT

Past research reveals best practices for creating SJTs. The first step is to develop the item content, typically done via critical incident interviews with SMEs. This is an essential step as SJTs that are created through such job analysis have significantly higher correlations with job performance. A critical incident interview requires the SMEs to think of examples of exceptionally good and exceptionally poor performance in a given job role. A recommended approach is the antecedent-behaviour-consequence (or A-B-C) format, which involves having the SMEs consider the context, the employee's behaviour, and the consequences of this behaviour, to get a full understanding of the situation.

SJTs can also be written to fit a conceptual model if there is existing job analysis information on the competencies required for the role or a theoretical model to draw on (MacCann & Roberts, 2008). Indeed, Campion et al. (2014) recommend including a more theoretical approach to increase reliability and, in turn, validity. Test creators can use a combination of both — first developing a model either from job analysis or theory, and then holding critical incident interviews to fill out the situations that occur within each section of the model.

Response options can be created either by the test developer or through additional consultation with SMEs. There is a lack of research on which is better, but Weekley and colleagues (2006) suggest that SMEs are likely to have a better understanding of the situations that could be faced in the role and thus are more likely to suggest realistic responses. Employees with a range of levels of experience should be approached to ensure a range of effectiveness in the suggested responses.

Next, a scoring key must be created. This can be done either through asking highly rated employees to rate the effectiveness of the options created in the previous step, or by having a pilot group complete the SJT and weighting the responses on the basis of their correlations with another criterion variable, such as job performance.

Recommendation for creating SJTs: Use a structured process that draws on SME knowledge. We recommend the use of SMEs throughout the development process to ensure that the situations and responses created are realistic and job relevant. When creating the item content, if the client has quality job analysis information then this may be used to identify the key competencies for role performance. Then, SMEs should be consulted in critical incident interviews to create the situations for the assessment within the competency framework. For example, if 'customer service' was identified as a key competency for the role, SMEs would be asked to think of really good and really poor performance relating to customer service. SMEs should be consulted also to create the content for the response options.

For creating the scoring key, we recommend an empirical approach if the client is willing to put a large pilot sample through the newly developed SJT and to provide a measure of job performance for the pilot participants to

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

allow evaluation of the scoring key. If this is not possible, an alternative is to have highly rated employees select the best and worst options.

Limitations of This Research

One limitation of this review is that we consider only three specific areas relating to SJTs. There are many more areas that could be reviewed to give a broader picture of SJTs and to provide other recommendations for practice. However, these issues may be investigated once the research field is more developed. Future questions include: 'Are SJTs more useful for some types of competencies?' and 'What factors affect how well SJTs predict job performance?' Also, similar to assessment centres, a fundamental question is: 'What exactly do SJTs measure?' (Chan, 2006).

An additional limitation relates to the differences between animation-based SJTs and video-based SJTs. We grouped video-based and animation-based SJTs under the heading of multimedia-based SJTs, yet the two are not identical. There was significantly less research available on the use of animated SJTs, which is not surprising as it is a newer area of testing. Future research could include questions such as 'Are animated characters able to portray emotions equally well as video characters?' and 'Are there differences in candidate perceptions of and reactions to animation-based compared to video-based SJTs?'

A third limitation is the lack of information in Australasian contexts. None of the articles reviewed were from Australasia, and research in this context is an important next step. Campion et al. (2014) note that context-free measures such as personality and cognitive ability show validity generalisation, hence this may be the case for SJTs. However, research to date has not examined the generalisability of the same SJT items and scoring keys across different contexts.

Implications for Minorities

Australia, New Zealand, and other countries in the Asia-Pacific region have diverse populations, including in some cases large numbers of immigrants. Moreover, in New Zealand it is a legal obligation to ensure that procedures are equitable for Māori. Some potential implications for non-White candidates can be drawn from the content reviewed. In the United States, comparisons of SJTs for Black and White students showed that video-based SJTs had lower adverse impact levels than text-based SJTs (Chan & Schmitt, 1997). Chan and Schmitt (1997) suggested this was due to the lower required levels of reading comprehension needed when there is no requirement to read text, but instead to watch and comprehend the situation. Thus, multimedia-based SJTs may also have positive implications in terms of reduced adverse impact for groups who tend to do less well on written ability tests, such as immigrants with English as a second language and also Māori who tend to score lower than population norms on tests of vocabulary (although higher than population norms for visuospatial tests of block design; Ogden & McFarlane-Nathan, 1997).

An additional consideration for countries with diverse ethnic populations is that multimedia-based SJTs, whether using real or animated people, should include people from different cultural backgrounds, with different body types, (dis)abilities, and of different ages. In line with this, research in other selection contexts has found that minority candidates respond favourably to representations of diversity during recruitment, with no adverse reaction from majority candidates (Avery, 2003).

Care should also be taken to ensure that different cultural groups are not discriminated against based on how they respond. For example, responses need to be carefully screened to ensure that both correct and incorrect answers are equally likely to be chosen by candidates regardless of their background, ethnicity, gender and age. Exploring this issue further, one can imagine that a response that involves collaboration with colleagues might be seen as more appropriate by candidates from collective cultures, or that deference to authority could be seen as more appropriate by those from high power distance cultures (Hofstede, 1999). One method to ensure fairness is to include a variety of people from different backgrounds in developing the response options and the scoring key.

Identification of Ethical Issues

As outlined above, different assessments can have different levels of adverse impact on ethnic groups. Although there is some evidence that multimedia-based SJTs appear to reduce adverse impact, with average scores having smaller differences between ethnic groups, this should be verified further by organisations when they put an SJT in place. An SJT should be used as part of a robust selection procedure and, as with any psychometric assessment, results should be considered in conjunction with other sources of information.

Furthermore, SJTs give organisations a chance to represent a variety of diverse people in their testing, particularly when using multimedia-based SJTs. Care should be taken to ensure that this is done respectfully and not in a stereotypical or insensitive manner.

Since SJTs may not look like a typical psychometric assessment, organisations using SJTs must be informed that the same laws apply to them and what these laws are. For example, in New Zealand, candidates must be informed of the purpose of the testing and what their results will be used for. Organisations must treat responses as personal information and look after these according to the processes outlined in New Zealand's *Privacy Act 1993*.

Finally, SJTs are likely to need ongoing validation. Among the various assessment methods available, SJTs are among those that are most able to leverage technological advances, such as gamification (DuVernet & Popp, 2014), compared with long-standing assessment methods

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

such as interviews which, while they can be implemented across different technology platforms (telephone, video), remain fundamentally the same. This means that SJT validity will need to be revisited with every technological update. Further to this, Campion et al. (2014) raise the issue of validity generalisation. Research is needed to assess whether an SJT developed for one setting could be used in an identical manner in a different setting and show similar validity. For practitioners hoping to leverage their intellectual property in developing SJTs, this issue of validity generalisation is critical. Moreover, jobs are continually changing and therefore SJT content may need to be changed to match both the reality of the work context (i.e., simulation fidelity) and the capabilities required for good job performance. In summary, validity is a constantly evolving characteristic of assessment, and therefore all assessments require ongoing validation (Messick, 1995), and this is especially the case for SJTs.

Conclusion

SJTs provide a contemporary, attractive and valid way to select candidates. Our findings reveal that multimedia presentation and *should* response options are optimal, and that the optimal development of SJTs is via job analysis procedures conducted with SMEs, although theoretical models may provide a useful adjunct. We hope that our review encourages more practitioners to develop and use SJTs, supports those in academic settings to understand some key issues for SJTs, and stimulates further research into the many remaining questions around SJT best practice.

References

Note: All book chapters and journal articles used for the literature review in this article have been marked with an asterisk.

Anderson, N., Herriot, P., & Hodgkinson, G.P. (2001). The practitioner-researcher divide in Industrial, Work and Organizational (IWO) psychology: Where are we now, and where do we go from here? *Journal of Occupational and Organizational Psychology*, *74*, 391–411. doi:10.1348/096317901167451

Anderson, N., Salgado, J.F., & Hülsheger, U.R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, *18*, 291–304. doi:10.1111/j.1468-2389.2010.00512.x

Arthur Jr, W., Glaze, R.M., Jarrett, S.M., White, C.D., Schurig, I., & Taylor, J.E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, *99*, 535–545. doi:10.1037/a0035788

Avery, D.R. (2003). Reactions to diversity in recruitment advertising — Are differences black and white? *Journal of Applied Psychology*, *88*, 672–679. doi:10.1037/0021-9010.88.4.672

Bauer, T.N., & Truxillo, D.M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J.A. Weekley & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 233–252). Mahwah, NJ: Erlbaum.

Briner, R., & Rousseau, D.M. (2011). Evidence-based I/O Psychology: Not there yet. *Industrial and Organizational Psychology*, *4*, 3–22. doi:1754-9426/11

*Brooks, M.E., & Highhouse, S. (2006). Can good judgment be measured? In J.A. Weekley & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 39–57). Mahwah, NJ: Erlbaum.

Campion, M.C., Ployhart, R.E., & MacKenzie Jr., W.I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, *27*, 283–310. doi:10.1080/08959285.2014.929693

Chan, D. (2006). Interactive effects of situational judgment effectiveness and proactive personality on work perceptions and work outcomes. *Journal of Applied Psychology*, *91*, 475–481. doi:10.1037/0021-9010.91.2.475

*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. *Journal of Applied Psychology*, *82*, 143–159. doi:10.1037/0021-9010.82.1.143

*Christian, M.S., Edwards, B.D., & Bradley, J.C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*, 83–117. doi:10.1111/j.1744-6570.2009.01163.x

DuVernet, A.M., & Popp, E. (2014). Gamification of workplace practices. *The Industrial-Organizational Psychologist*, *52*, 39–42.

Grant, A.M., & Pollock, T.G. (2011). Publishing in AMJ — Part 3: Setting the hook. *Academy of Management Journal*, *54*(5), 873–879. doi: 10.5465/amj.2011.4000

Griffin, D.W., Dunning, D., & Ross, L. (1990). The role of construal processes in overconfident predictions about the self and others. *Journal of Personality and Social Psychology*, *59*, 1128–1139. doi:10.1037/0022-3514.59.6.1128

Hofstede, G. (1999). Problems remain, but theories will change: The universal and the specific in 21st-century global management. *Organizational Dynamics*, *28*, 34–44. doi:10.1016/S0090-2616(00)80005-4

*Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*, 426–441. doi:10.1108/00483480810877598

*Lievens, F., & Sackett, P.R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*, 1181–1188. doi:10.1037/0021-9010.91.5.1181

*Lievens, F., Sackett, P.R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, *94*, 1095–1101. doi:10.1037/a0014628

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, *116*, 75–98. doi:10.1006/obhd.1996.0028

Loewenstein, G., & Schkade, D. (1999). Wouldn't it be nice? Predicting future feelings. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 85–105). New York: Sage.

MacCann, C., & Roberts, R.D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, *8*, 540–551. doi:10.1037/a0012746

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press

*McDaniel, M.A., Hartman, N.S., Whetzel, D.L., & Grubb, W.L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60*, 63–91. doi:10.1111/j.1744-6570.2007.00065

McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., & Campion, M.A. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730–740. doi:10.1037//0021-9010.86.4.730

McDaniel, M.A., & Nguyen, N.T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*, 103–113. doi:10.1111/1468-2389.0016

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003- 066X.50.9.741.

Motowidlo, S.J., Crook, A.E., Kell, H.J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology*, *24*, 281–288. doi:10.1007/s10869-009-9106-4

*Nguyen, N.T., Biderman, M.D., & McDaniel, M.A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment*, *13*, 250–260. doi:10.1111/j.1468-2389.2005.00322.x

*Nguyen, N.T., & McDaniel, M.A. (2003). Response instructions and racial differences in a situational judgment test. *Applied H.R.M. Research*, *8*, 33–44.

Ogden, J.A., & MacFarlane-Nathan, G. (1997). Cultural bias in the neuropsychological assessment of young Māori men. *New Zealand Journal of Psychology*, *26*, 2–12.

*Olson-Buchanan, J.B., & Drasgow, F. (2006). Multimedia situational judgment tests: The medium creates the message. In J.A. Weekley & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 253–278). Mahwah, NJ: Lawrence Erlbaum Associates.

*Ployhart, R.E., & Ehrhart, M.G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment*, *11*, 1–16. doi:10.1111/1468-2389.00222

Privacy Act of 1993. §§ 1–12. Retrieved from http://www.legislation.govt.nz/act/public/1993/0028/latest/DLM296639.html

*Richman-Hirsh, W.L., Olson-Buchanan, J.B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology*, *85*, 880–887. doi:10.1037/0021-9010.85.6.880

Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K., & Stoffey, R.W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, *46*, 49–76. doi:10.1111/j.1744-6570.1993.tb00867.x

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? *International Journal of Selection and Assessment*, *22*, 225–239. doi:10.1111/ijsa.12072

*Weekley, J.A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology*, *50*, 25–49. doi:10.1111/j.1744-6570.1997.tb00899.x

*Weekley, J.A., Ployhart, R.E., & Holtz, B.C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J.A. Weekley & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157–182). Mahwah, NJ: Lawrence Erlbaum Associates.

https://doi.org/10.1017/orp.2015.6 Published online by Cambridge University Press