# 11

# Situational Judgement Tests for Selection

Jan Corstjens, Filip Lievens and Stefan Krumm

## Introduction

When situational judgement tests (SJTs) began to regain popularity among the scientific community in the 1990s, there was an implicit notion that they captured context-dependent knowledge. In fact, the term 'situational judgement' carries the connotation of test-takers' responses being more effective when they consider the specifics of the situation. In recent years another perspective has emerged, which views SJTs as capturing relatively context-independent knowledge (or general domain knowledge; Motowidlo, Crook, Kell & Naemi, 2009; Motowidlo, Hooper & Jackson, 2006a). Although SJTs and their items will often fall somewhere between these two perspectives, we posit in this chapter that it might be useful to distinguish between them. So far, there has been no review of the SJT literature in terms of these two approaches. This is understandable, as over the years the two perspectives have emerged alongside each other. Therefore, the aim of this chapter is to review SJT research according to these two approaches.

The chapter is structured as follows. We start by presenting the traditional contextualized perspective underlying SJTs. We review the underlying theory, the developmental stages and the research evidence regarding this perspective (e.g., reliability, criterion-related validity, construct-related validity, subgroup differences, applicant reactions). We end our discussion of the contextualized perspective by homing in on new trends. Next, we present the general domain knowledge perspective, thereby following exactly the same structure as for the contextualized perspective. We end this chapter by presenting directions for future research and by giving recommendations for HR practices.

*The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention*,
First Edition. Edited by Harold W. Goldstein, Elaine D. Pulakos, Jonathan Passmore and Carla Semedo.

## Contextualized SJTs

### The underlying rationale and theory

Simulations represent contextualized selection procedures that psychologically and/or physically mimic key aspects of the job (Lievens & De Soete, 2012). In accordance with this definition, contextualized SJTs aim to confront applicants with a set of situations similar to those they might encounter on the job and elicit their procedural knowledge about how to respond to these stimuli. Like other simulations such as assessment centre exercises or work samples, context-specific SJTs rest on the notions of point-to-point correspondence with the criterion (future job situations) and behavioural consistency (Bruk-Lee, Drew & Hawkes, 2014; Lievens & De Soete, 2012). Behavioural consistency denotes that candidates' performance on a selection test will be consistent with their future job performance (Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968). To this end, simulations should ideally be constructed in such a way that there is a high degree of correspondence between the conditions in the simulation and those in the actual job context and tasks. Assessment centre exercises, for example, mimic actual job situations and generate behavioural samples and hence are referred to as high-fidelity simulations (Thornton & Rupp, 2006). Fidelity refers to the degree to which the simulation authentically reflects the targeted job in terms of both stimuli and responses (Motowidlo, Dunnette & Carter, 1990). To reduce development and administration costs of such simulations, most SJTs adopt a low-fidelity format in simulating the situations and responses. That is, SJTs typically present written (or video-based) descriptions of job-related situations and require a response to them by opting for an alternative from a list of multiple-choice responses (McDaniel, Hartman, Whetzel & Grubb, 2007; Weekley, Ployhart & Holtz, 2006).

Notably, situation descriptions are key to SJTs when viewed from the contextualized perspective because they simulate job contexts, guide candidates' situation perception and subsequent response selection and render responses more or less effective. Thus, the situation descriptions in SJTs aim to provide sufficient contextualization so that candidates can imagine the situation and make well-thought-out judgements about how they would or should behave according to the situational demands depicted (Richman-Hirsch, Olson-Buchanan & Drasgow, 2000). So, this view assumes that test-takers' behavioural response selection is contingent on how they perceive and construe the stimuli (job-related situations), which aligns well with interactionist theories that consider behaviour to be a function of both the person's traits and the person's perception of the situation (Campion & Ployhart, 2013; Mischel & Shoda, 1995). Each situation conveys specific cues, which are interpreted by each test-taker. The person's interpretation of the cues is guided by previous experiences in similar situations and determines the response selection believed to be appropriate. Without this context, it is assumed the test-taker is left in the dark as to what the appropriate response should be and might lack sufficient information to solve the item.

### Developmental stages

The typical steps involved in developing contextualized SJTs are threefold (Lievens, Peeters & Schollaert, 2008; Motowidlo et al., 1990). The first stage concerns the development of item stems or situations to be presented in the SJT. The second stage involves the collection of response options from subject matter experts (SMEs), the choice of response instructions and of the response format. The third and final stage targets the development of the scoring key.

*Stage 1: Item stems* To gather the item stems or situations presented in the SJT, a job analysis is usually conducted. During this job analysis, SMEs are asked to generate critical incidents (Flanagan, 1954), which means that they are asked to recall examples of situations in which exceptionally good or exceptionally poor performance was demonstrated. The test developer often prompts the SMEs with the goal of collecting information about all the content domains and constructs deemed to be important for the job. The selected SMEs are typically incumbents, supervisors, managers or a mix of these sources of information. Alternatively, archival sources and even customers might serve as a source of information (Weekley et al., 2006). The critical incidents obtained are then sorted and checked for redundancy and level of specificity. The surviving incidents then serve to write item stems or descriptions of job-related situations. As an alternative to this inductive method of gathering critical incidents, a deductive method can be followed. In this strategy, the item stem content is derived from theoretical models (e.g., a model of conflict management).

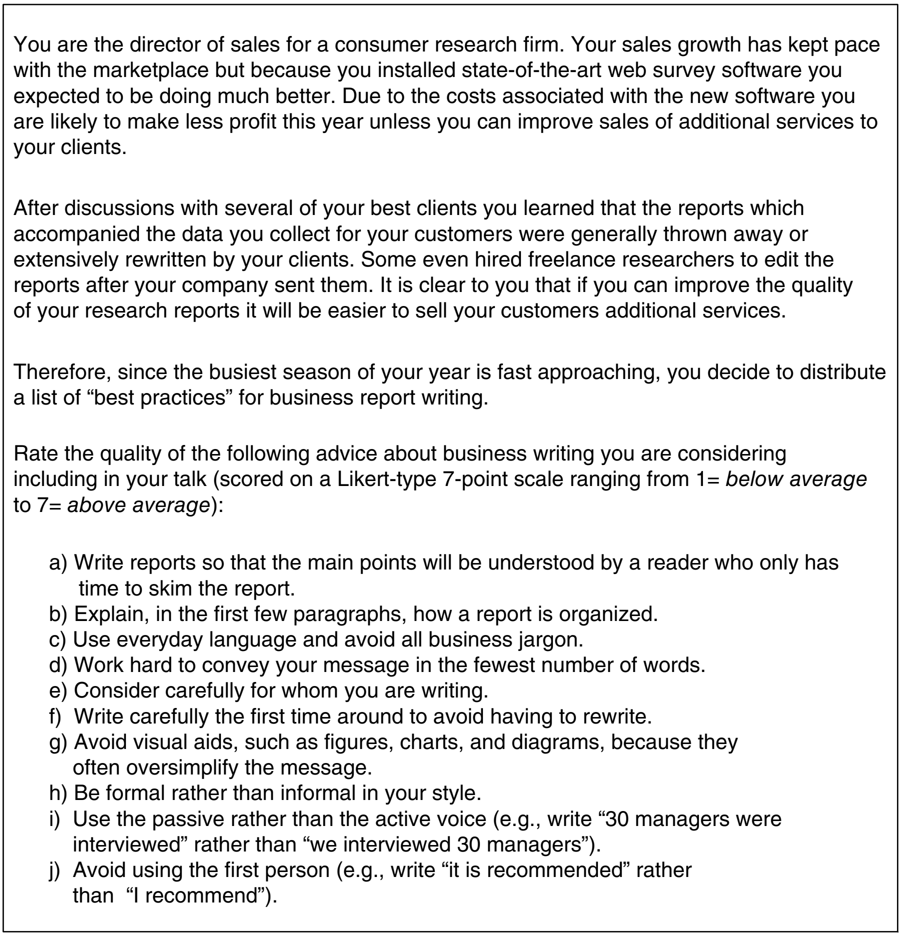
*Stage 2: Response options, response instructions, and response format* After developing the situation descriptions, another group of SMEs is asked to generate response options they believe to be (in-)effective reactions to the situations. To obtain a wider range of response options with different levels of effectiveness, the test developer might also ask inexperienced workers to generate responses. The test developer then decides which options to retain, usually by choosing a mix of response options that are differentially effective in each situation. There are no general rules regarding the number of response options to retain. The majority of SJT items include 4 or 5 response options, even though SJT items with up to 10 response options also exist (e.g., the Tacit Knowledge Inventory; Wagner & Sternberg, 1991).

In the next stage, the test developer decides on the response instructions. This is not a trivial choice because the response instruction format affects the construct saturation of the SJT (McDaniel et al., 2007). One of two formats of response instructions is usually chosen: behavioural tendency instructions or knowledge instructions (McDaniel & Nguyen, 2001). Behavioural tendency instructions ask respondents what they would do in the given situation, whereas knowledge instructions ask respondents what they should do in the situation; in other words, they ask respondents to identify the best response to a given situation.

Test developers also make a choice about the response format to be employed. Generally, three response formats can be distinguished. Respondents are asked to select the best/worst response options, rank the response options from most to least effective or rate the response options on Likert-type scales. Arthur, Glaze, Jarrett, White, Schurig and Taylor (2014) comparatively evaluated these three common response formats by varying them while keeping the rest of the SJT design and content constant. The rate response format evidenced higher construct-related validity, lower levels of subgroup differences and increased reliability over the other two. A drawback of the rate response format, however, was its higher susceptibility to response distortion.

*Stage 3: Scoring key* After situations, response options, response format and instructions have been developed, the test requires a scoring key. Here, four different methods can be delineated. The rational method involves asking a group of SMEs to score the response options on (in-)effectiveness. Scores with acceptable inter-rater agreement (e.g., $\geq 0.60$) are retained for the test. The second is the empirical method which involves quantifying endorsements of correct response options gathered from a large sample of lay people instead of SMEs. For instance, options that are chosen to be correct by over 25% of the sample are retained for the test. Although notably different in approach, researchers have found no differences between these two scoring keys in terms of validity (e.g., Weekley & Jones, 1999). Combining the rational and empirical method is a third approach that can be followed. An example of this hybrid approach is retaining an empirical key only after SMEs have agreed on it. The final and least frequently followed method involves the development of a scoring key with answer options that reflect effective performance according to a chosen theoretical framework (e.g., leadership theories) scored as correct (Weekley et al., 2006).

In Figure 11.1 we present an example of a contextualized SJT item that was taken from the Tacit Knowledge Inventory for Managers (Wagner & Sternberg, 1991). Although not strictly called an SJT by the developers, the test is similar to the format and content of a typical SJT (McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001).



You are the director of sales for a consumer research firm. Your sales growth has kept pace with the marketplace but because you installed state-of-the-art web survey software you expected to be doing much better. Due to the costs associated with the new software you are likely to make less profit this year unless you can improve sales of additional services to your clients.

After discussions with several of your best clients you learned that the reports which accompanied the data you collect for your customers were generally thrown away or extensively rewritten by your clients. Some even hired freelance researchers to edit the reports after your company sent them. It is clear to you that if you can improve the quality of your research reports it will be easier to sell your customers additional services.

Therefore, since the busiest season of your year is fast approaching, you decide to distribute a list of “best practices” for business report writing.

Rate the quality of the following advice about business writing you are considering including in your talk (scored on a Likert-type 7-point scale ranging from 1= *below average* to 7= *above average*):

a) Write reports so that the main points will be understood by a reader who only has time to skim the report.
b) Explain, in the first few paragraphs, how a report is organized.
c) Use everyday language and avoid all business jargon.
d) Work hard to convey your message in the fewest number of words.
e) Consider carefully for whom you are writing.
f) Write carefully the first time around to avoid having to rewrite.
g) Avoid visual aids, such as figures, charts, and diagrams, because they often oversimplify the message.
h) Be formal rather than informal in your style.
i) Use the passive rather than the active voice (e.g., write “30 managers were interviewed” rather than “we interviewed 30 managers”).
j) Avoid using the first person (e.g., write “it is recommended” rather than “I recommend”).

**Figure 11.1** Example of a contextualized SJT item. *Source*: Wagner & Sternberg (1991). Reproduced with permission of Robert J. Sternberg.

## Overview of Prior Research

Research on SJTs has mushroomed following their reintroduction in the academic literature by Motowidlo and colleagues (1990). The vast majority of research evidence on SJTs pertains to the contextualized view as the traditional perspective on SJTs. In this section, we review such research evidence concerning reliability, criterion-related and incremental validity, construct-related validity, subgroup differences, applicant reactions, faking, retest and coaching effects. Whenever meta-analytic findings are available, we refer to them.

### Reliability

Several meta-analyses have integrated internal consistency reliability coefficients that have been reported in the SJT literature. The mean α values reported in these meta-analyses ranged from 0.46 to 0.68 (Campion, Ployhart & MacKenzie, 2014; Catano, Brochu & Lamerson, 2012; Kasten & Freund, 2015). The reason for the moderate internal consistency reliability coefficients is the fact that SJTs are created on the basis of job situations that require the expression of a combination of different constructs, which results in heterogeneous test items *and* response options. Evidence for item heterogeneity comes from factor analytic investigations of SJTs that reveal no clear factor structure in the items (Schmitt & Chan, 2006).

As internal consistency is not a suitable reliability estimate for a measurement method that has heterogeneous items (Osburn, 2000), other types of reliability estimates, such as test–retest reliability and alternative form reliability, have been proposed in the literature (Lievens et al., 2008; Whetzel & McDaniel, 2009). Studies examining test–retest reliability are scarce but they tend to report considerably higher estimates. For instance, Catano and colleagues (2012) reported two SJT test–retest coefficients of $r$ = 0.82 and $r$ =0.66, respectively. Studies examining alternative form reliability coefficients are even scarcer because of the difficulty in developing alternative form SJTs that capture the same constructs when these constructs are often not clearly distinguishable to begin with. Notwithstanding this, Clause, Mullins, Nee, Pulakos and Schmitt (1998) reported alternative test reliability estimates ranging from $r$ = 0.70 to $r$ = 0.77 when they adopted a rigorous item cloning method for constructing alternative SJT forms (see also Lievens & Sackett, 2007). So, SJTs have generally been found to be sufficiently reliable measurement instruments, provided that appropriate reliability estimates are used.

### Criterion-related and incremental validity

Much data have accumulated over the years supporting the relation between SJTs and job performance. McDaniel and colleagues conducted two meta-analyses (McDaniel et al., 2001, 2007) and reported corrected estimated population correlations of 0.26 and 0.34, respectively (uncorrected correlations 0.20 and 0.26). The more recent analysis included data on over 24,000 respondents. The criterion used in most studies is a composite score of job performance ratings. However, as evidenced by Christian and colleagues' meta-analytic findings, criterion-related validity can increase when predictor and criterion are more carefully matched (Christian, Edwards & Bradley, 2010). These authors divided the job performance criterion into three facets: task performance (i.e., job-specific skills), contextual performance (i.e., soft skills and job dedication), and managerial performance (i.e., management skills). SJTs were then sorted into a typology of construct domains. The authors hypothesized that criterion-related validity would increase if particular criterion facets were closely matched with the content domains of the SJTs (e.g., contextual performance predicted by SJTs from the domains of interpersonal and teamwork skills). Overall, the authors found support for their content-based matching approach: relatively homogeneous SJTs saturated with a particular construct domain evidenced higher criterion-related validity with the criterion component they were designed to predict than heterogeneous composite SJTs.

In addition to moderation by criterion facet (a content-based moderator), the criterion-related validity of an SJT can be influenced by method-based moderators. We highlight three moderators identified in the literature relating to 1) test development procedure, 2) item stem format and 3) test delivery format. Meta-analytic evidence established that SJTs yield higher validities ($r$ =0.38 vs. $r$ = 0.29) when they are based on a careful job analysis than when they are based on intuition or theory (McDaniel et al., 2001). A second moderator is the level of detail in the item stem; less detailed questions show a slightly larger validity than highly detailed questions ($r$ = 0.35 vs. $r$ = 0.33). This runs somewhat counter to the premise of contextualized SJTs that context and level of detail increase the criterion-related validity of the test scores. Third, the test delivery format has been found to differentially affect validity, with video-based SJTs showing higher levels of criterion-related validity for predicting interpersonal skills than the traditional paper-and-pencil format, with a corrected population correlation of 0.36 for video-based SJTs and 0.25 for paper-and-pencil formats (Christian et al., 2010). This finding supports the contextualized perspective of SJTs because contextual information (e.g., about environmental cues, nonverbal behaviour) seems to be necessary to adequately apply interpersonal skills.

An interesting strand of research concerns investigating the incremental validity of SJTs as compared to other predictors of performance. McDaniel and colleagues (2007) found that SJTs explained 6–7% additional variance above the Big Five personality factors and 3–5% additional variance above cognitive ability, depending on the type of response instruction (knowledge instructions vs. behavioural tendency instructions). Further, SJTs explained 1–2% of variance above both cognitive ability and the Big Five factor scores. More recently, SJTs as low-fidelity simulations have been contrasted with assessment centre exercises in a high-stakes selection context. Lievens and Patterson (2011) found that criterion-related validity was similar for both the SJT and assessment centre exercises. Subsequent incremental validity analyses revealed that the assessment centre exercises explained 3% additional variance in the criterion job performance over the SJT. However, subsequent path analysis showed that assessment centre performance only partially mediated the effect of procedural knowledge as measured by the SJT on job performance, indicating that scores obtained from these two types of simulations should not be viewed as redundant.

In sum, contextualized SJTs predict variance in job-related criteria to an extent that is comparable to other frequently used selection tools (see Schmidt & Hunter, 1998). Importantly, contextualized SJTs contribute incrementally above and beyond Big Five personality factors and general mental ability.

*Construct-related validity* For the same reason that makes it difficult to estimate internal consistency reliability of SJT scores, item heterogeneity makes it challenging to delineate which construct(s) are being measured by the SJT. Next to decisions pertaining to the actual test content, the method of measurement can also influence which constructs are being measured by SJTs. Concerning measurement method, McDaniel and colleagues (2007) obtained a differential pattern of construct-related validity coefficients when SJTs with knowledge instructions ('What should you do in a given situation?') were compared to SJTs with behavioural tendency instructions ('What would you do in a given situation?'). Correlations between SJTs with behavioural tendency instructions and three Big Five personality factors were higher than for SJTs with knowledge instructions (agreeableness 0.37 vs. 0.19, conscientiousness 0.34 vs. 0.24, and emotional stability 0.35 vs. 0.12, respectively). Conversely, SJTs with knowledge instructions correlated at a higher rate with measures of cognitive ability than SJTs with behavioural tendency instructions (0.35 vs. 0.19, respectively).

*Subgroup differences* Although SJTs generally result in smaller subgroup differences than cognitive ability tests, they are not absent in SJTs (Lievens et al., 2008). Whetzel, McDaniel and Nguyen (2008) meta-analytically investigated race and gender as two demographic variables that can lead to subgroup differences in SJT scores. Regarding gender, females in general performed slightly better than males ($d$ = 0.11). Concerning race, they found that Whites performed better than Blacks ($d$ = 0.38), Hispanics ($d$ = 0.24) and Asians ($d$ = 0.29). Subgroup differences were not invariant across all SJTs because several moderators have been found to influence the relation with SJT performance. Racial differences, for example, could be explained by the cognitive loading of the SJT. That is, SJTs that were more correlated with general mental ability resulted in larger racial differences than SJTs that were more correlated with personality constructs (Whetzel et al., 2008). Reduced racial differences were also observed when behavioural tendency instructions were used instead of knowledge instructions (differences between Whites and Blacks of $d$ = 0.39 for knowledge instructions and $d$ = 0.34 for behavioural tendency instructions; Whetzel et al., 2008), and when video-based SJTs were used ($d$ = 0.21 compared to a paper-and-pencil SJT, Chan & Schmitt, 1997). In contrast to racial differences, gender differences seemed to increase only when the personality loading of the SJT increased, thereby favouring women ($d$ = –0.37 and –0.49 as compared to men for conscientiousness and for agreeableness, respectively) and remained invariant when the cognitive loading increased (Whetzel et al., 2008).

Other than the cognitive loading of SJTs, McDaniel and colleagues (2011) suggested that more extreme response tendencies might also explain Black–White subgroup differences in SJT scores and proposed controlling for these response tendencies in SJT scoring. They administered SJTs with Likert-type scales in two concurrent designs and subsequently adjusted the scale scores for elevation and scatter (i.e., respondents' item means and deviations). Their strategies resulted in a reduction of Black–White mean score differences across the two measurement occasions, with effect sizes dropping from around half an *SD* ($d$ = 0.43–0.56) to about a third of an *SD* ($d$ = 0.29–0.36) for the standardized scores to less than a fifth of an *SD* for the dichotomous scoring ($d$ = 0.12–0.18). Roth, Bobko and Buster (2013) highlighted a caveat in this subgroup differences SJT research, namely that the studies have nearly always been conducted with concurrent designs (i.e., samples consisting of job incumbents and not applicants). A sole focus on concurrent designs could lead to range restriction attenuating the obtained effect sizes and thus to an underestimation of effect sizes in the population (see also Bobko & Roth, 2013). These authors argue that in order to reduce the potential issue of range restriction, subgroup differences should also be studied in samples of applicants who are assessed with the SJT at the earliest possible selection stage (and before any other measures have been deployed). In such applicant samples findings pointed towards Black–White differences of $d$ = 0.63 for SJTs that were mainly saturated with cognitive ability, $d$ = 0.29 for SJTs saturated with job knowledge and $d$ = 0.21 for SJTs that mainly tapped interpersonal skills. These results further confirm previous findings of racial differences increasing with the cognitive loading of the SJT.

*Applicant reactions* In general, research has demonstrated that applicants prefer selection tools they perceive as job-related, that provide opportunities to show their capabilities and that are interactive (e.g., face-to-face interviews) (Hausknecht, Day & Thomas, 2004;

Lievens & De Soete, 2012; Potosky, 2008). High-fidelity simulations typically contain many of these aspects. Several studies have shown that applicant reactions to low-fidelity SJTs also tend to be favourable, and even more so when fidelity is increased and interactivity is added. Chan and Schmitt (1997) showed that a video-based SJT received higher face validity ratings than a written SJT. Richman-Hirsch and colleagues (2000) found that interactive video-based formats were preferred to computerized and paper-and-pencil formats. In an interactive (branched or nonlinear) SJT, the test-taker's previous answer is taken into account and determines the way the situation develops. Kanning, Grewe, Hollenberg and Hadouch (2006) went a step further and varied not only stimulus fidelity (situation depicted in a video vs. written format), but also response fidelity (response options shown in a video vs. written format) and interactivity of SJTs. In line with the previously mentioned studies, applicants reacted more favourably towards interactive video-based formats, and in this case towards both the stimulus and the response format.

*Faking, retesting and coaching* Hooper, Cullen and Sackett (2006) compiled the research findings on faking and discovered that there was a lot of variation concerning the relation between faking and SJT performance: effect sizes ranged from $d = 0.08$ to 0.89 suggesting the presence of moderators. One such moderator proposed by the authors is the cognitive or *g* loading of the items. Although based on just a handful of studies, the trend is that SJTs with higher cognitive loadings are less easy to fake (Hooper et al., 2006; Peeters & Lievens, 2005). Similarly, the degree of faking can vary depending on the response instructions, with knowledge instructions being less easy to fake than behavioural tendency instructions (Nguyen, Biderman & McDaniel, 2005).

As SJTs are often part of large-scale, high-stakes selection programmes, it is also important to examine whether retest and coaching effects influence test scores and their psychometric properties. Concerning retest or practice effects, Lievens, Buyse and Sackett (2005) reported effects of $d = 0.29$ (0.49 after controlling for measurement error). A similar result was found by Dunlop, Morrison and Cordery (2011), who found an effect size of $d = 0.20$. Importantly, in both studies retest effects were found to be smaller for SJTs in comparison to cognitive ability tests. Dunlop and colleagues further noticed that practice effects decreased at a third measurement occasion for both the SJT and the cognitive ability tests. As far as coaching is concerned, only two studies have tackled this issue to date. Cullen, Sackett and Lievens (2006) investigated the coachability of two college admission SJTs and found that coaching increased the scores on one of the SJTs ($d = 0.24$) but not on the other. In contrast to Cullen and colleagues' study which took place in a laboratory setting, Lievens, Buyse, Sackett and Connelly (2012) investigated coaching on SJT scores in a high-stakes setting. Moreover, the latter study included pretest and propensity score covariates to control for self-selection in order to reduce the non-equivalence of the groups. Using this more sophisticated analysis, they found that coaching raised SJT scores with 0.53 SDs. Finally, a recent study (Stemig, Sackett & Lievens, 2015) found that organizationally endorsed coaching (i.e., coaching provided by the organization rather than commercial coaching) also enabled people to raise their SJT scores, but did not reduce the criterion-related validities of the SJT scores.

In sum, contextualized SJTs seem to be less prone to faking and retest effects than other selection methods. Such effects may be further reduced by using knowledge-based response instructions and developing SJTs with higher *g* loadings. Coaching effects can be reduced by enabling all candidates to practice on SJTs in advance of high-stakes assessments.

## Contextualized SJTs: Implications and Trends

The contextualized perspective of SJTs has important implications for SJT design as it encourages test developers to increase the SJT situations' level of contextualization and fidelity. Over the years, various innovations have been proposed as alternatives to classic paper-and-pencil SJTs. These innovations have focused on increasing the realism of the situation depicted (i.e., stimulus fidelity) or the realism of the manner in which applicants are able to respond (i.e., response fidelity).

A well-known example of increasing stimulus fidelity consists of using video-based or multimedia formats instead of written scenarios. Recently, advances in terms of both 3D animation and motion-capture techniques have been employed by SJT developers as a way to increase stimulus fidelity (Weekley, Hawkes, Guenole & Ployhart, 2015). Companies that make use of technologically advanced selection tests also look more appealing to the contemporary, tech-savvy generation of gamers and internet users (Fetzer & Tuzinski, 2014). The use of 3D animation has several advantages over video-based SJTs. First, the costs involved in hiring actors and film crews are reduced since only voice actors and software programmers are required. Second, 3D animation can be more flexible than video-based SJTs because in the latter some situations cannot be filmed due to cost concerns and consequently have to be excluded (e.g., a factory fire). Third, 3D animations allow customization in different contexts and cultures. For example, with a little bit of programming one can change the gender and ethnic background of the characters depicted (Fetzer, Tuzinski & Freeman, 2010).

Motion-capture techniques are another recent development. They make use of live actors whose movements and facial expressions are registered by markers placed on the body and face. The computer registers the signals sent from these markers and the actors' movements and expressions are then digitally converted into the software environment. Motion-capture techniques make programming of movements themselves redundant and therefore require less time and effort from programmers (Fetzer et al., 2010). Although these technologies are intuitively appealing, research has not been able to catch up with these fast-paced developments and comparative research with more traditional SJTs has been lacking up to this point.

Another way to increase realism is to enhance the response fidelity of an SJT. Instead of giving applicants descriptions of possible behavioural response options, the test can be constructed to capture candidates' actual behavioural responses to the situations (e.g., via a webcam; see Oostrom, Born, Serlie & van der Molen, 2010). In this case, SJT responses resemble the behavioural responses typically demonstrated in assessment centre exercises and allow the measurement of (non-)verbal and paralingual communication and expressions of emotions. In occupations where communication skills are important, assessment of such responses might increase the SJT's point-to-point correspondence with the criterion and result in higher validity for predicting job performance than responses captured via multiple-choice formats. Lievens, De Corte and Westerveld (2015) compared two multimedia SJTs (one with written constructed responses, one with webcam-captured responses) for predicting police officer job performance. They found evidence of significant incremental validity (2.8–8.3% of additional explained variance) and higher media richness perceptions for the open-ended format that captured candidates' behaviour via webcam. Investing in response-gathering technologies such as webcam SJTs therefore seems warranted because research shows increases in validity (Oostrom et al., 2010), positive candidate reactions (Bruk-Lee et al., 2014) and decreases in test score subgroup differences because of their lower cognitive loading (e.g., De Soete, Lievens, Oostrom & Westerveld, 2013).

Although increasing the fidelity of the situation and the response format of SJTs undeniably makes the test more realistic, SJTs still proceed through situations in linear fashion. In other words, once a response option has been adopted or expressed, the tests proceed to the next situation. Another way to make SJTs more realistic and contextualized is then to present situations that depict the consequences of the choices that were made in the initial situation and assess how the candidate responds to these new developments. This can be achieved through item branching where subordinate situation stems are activated depending on the response that has been chosen or made in the 'mother' stem (Weekley et al., 2015). Technological advances in the gaming industry have inspired some selection test developers to create virtual sandbox environments that allow the implementation of such item branching. These adaptive simulations or serious games could very well become the future of SJTs and selection tests in general. However, the more these environments become unscripted and unstructured, the harder it becomes to accurately assess constructs and/or traits deemed to be important for the job (Fetzer & Tuzinski, 2014).

## General Domain Knowledge SJTs

### Underlying rationale and theory

In the past few years, an alternative paradigm has emerged which views SJTs as measures of general domain knowledge that is seen as more context-independent. In a series of papers, Motowidlo and colleagues (Motowidlo & Beier, 2010; Motowidlo, Hooper & Jackson, 2006a,b) provided the conceptual foundation for this perspective. According to these researchers, general domain knowledge pertains to general rules about the utility of behavioural acts across a wide range of situations in a specific domain. The more general this knowledge is, the more it is context-independent and the more it is broadly applicable across a wide range of situations. Importantly, general domain knowledge is not acquired from specific job experiences. Rather, general domain knowledge reflects fundamental socialization processes (parents, schooling, etc.) and personal dispositions. That is why this general domain knowledge is also referred to as implicit trait policies (ITPs; Motowidlo & Beier, 2010; Motowidlo et al., 2006b), which are inherent beliefs about the general effectiveness of actions that express traits to varying degrees. In addition, people might have learned exceptions in situations where their inherent trait expressions were not as effective and as a result had to update and modify their ITPs (Motowidlo & Beier, 2010; Motowidlo et al., 2006a). Motowidlo and Beier (2010) further refined their theory of knowledge determinants underlying SJT performance by distinguishing more explicitly between general domain knowledge and specific job knowledge as the two components making up procedural knowledge as captured by an SJT. They first demonstrated that their SJT from 1990 (which was taken to be a contextualized measure) mainly captures general domain knowledge because two scoring keys with effectiveness ratings obtained from both novices and experts largely overlapped and both were significantly related to job performance. Second, the expert key showed incremental variance (5.2%) over the novice key, indicating that while for the most part the SJT captured general domain knowledge, there was still a component of procedural knowledge that could not be solved on the basis of general domain knowledge alone. According to the authors, these expert residual scores reflect specific job knowledge, which is mostly acquired in the job or family of jobs that the SJT is targeting. Cognitive ability and personality are posited as antecedents to these two forms of procedural knowledge as captured by the SJT. The relationship between ability and procedural knowledge is based on the mechanism of one's capacity to learn. Conversely, the relationship between personality traits and ITPs is grounded by the mechanism of dispositional fit. That is, personality traits interact with traits expressed by the different actions put forward in the SJT items in such a way that people who possess high levels of the trait expressed by the action believe that their action is truly more effective than people who have a lower standing on the trait. For instance, when judging the effectiveness of behaviours described in the response options of an SJT, individuals high on the trait of agreeableness will favour those response options that express higher levels of agreeableness more than individuals low in agreeableness (Motowidlo et al., 2006b).

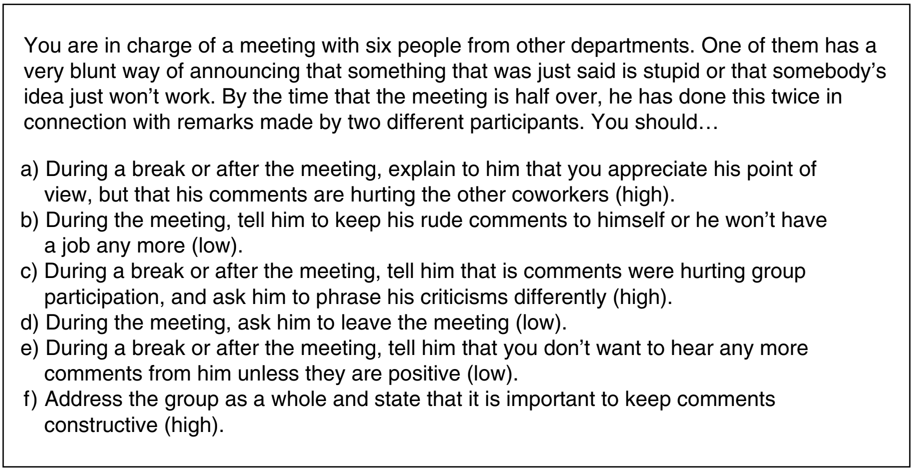
## Developmental stages

As is the case for contextualized SJTs, the development process of general domain knowledge SJTs can be categorized into three main steps. However, as compared to contextualized SJTs, each of these steps differs when applied to the measurement of general domain knowledge.

*Step 1: Item stems* According to the general domain knowledge perspective, each stem needs to be designed in such a way that the stem activates the constructs reflected in the response options, thereby allowing people to show their varying levels of procedural knowledge about these targeted constructs. This means that the test designer should adopt a strategy to develop item situations (item stems) on the basis of theoretical frameworks or taxonomies so that these situations can activate specific behaviour related to the targeted traits or compound traits (or competences; Motowidlo et al., 2006a; Patterson, Ferguson, Norfolk & Lane, 2005). In other words, under the domain-general design scheme, the development of item stems mainly follows a deductive approach rather than an inductive approach. However, to guarantee the job-relatedness of the situations, it is sometimes important (though not absolutely necessary) to 'beef up' these situations with information from critical incident interviews or workshops. In any case, test developers are advised to keep the situation descriptions quite generic. An SJT measuring general domain knowledge requires just enough job-specific contextualization to make the SJT face valid and job-related.

*Step 2: Response options, response instructions, and response format* Collecting response options for general domain knowledge SJTs does not require a group of experienced SMEs with considerable job-specific knowledge about the domains to be tapped by the SJT, because the response options are intended to tap expressions of general domain knowledge. For instance, a sample of novices or industrial and organizational psychology students (because they have knowledge of traits and trait expressions) can be instructed to generate response options by asking them what they think would be the best way to handle the situation presented in each item stem (Motowidlo et al., 2006a). The test developer then edits these responses. A sample of 5–10 response options are then typically retained per item stem, with an equal number of response options that express high levels of the trait and low levels of the trait (effective vs. ineffective options).

To assess general domain knowledge, a knowledge-based response instruction format ('What should you do?') seems to be most appropriate. Applicants should be instructed to give effectiveness ratings for each option. In that case, the response format is typically a Likert-type scale rating format, although pick best/worst and rank order formats are also possible.



You are in charge of a meeting with six people from other departments. One of them has a very blunt way of announcing that something that was just said is stupid or that somebody's idea just won't work. By the time that the meeting is half over, he has done this twice in connection with remarks made by two different participants. You should…

a) During a break or after the meeting, explain to him that you appreciate his point of view, but that his comments are hurting the other coworkers (high).
b) During the meeting, tell him to keep his rude comments to himself or he won't have a job any more (low).
c) During a break or after the meeting, tell him that is comments were hurting group participation, and ask him to phrase his criticisms differently (high).
d) During the meeting, ask him to leave the meeting (low).
e) During a break or after the meeting, tell him that you don't want to hear any more comments from him unless they are positive (low).
f) Address the group as a whole and state that it is important to keep comments constructive (high).

**Figure 11.2** Example of general domain knowledge SJT item (related to agreeableness). *Source*: Motowidlo, Hooper & Jackson (2006a).

*Step 3: Scoring key* SMEs with extensive knowledge about the varying personality trait expressions in the response options are required to develop the scoring key. For the measurement of the personality trait conscientiousness, for example, personality psychologists or doctoral students in the domain of personality psychology could be approached to rate the response options. To this end, Likert-type scales can be used with verbal labels indicating the level of the trait expressed by the response option (e.g., 1 = very introverted to 7 = very extraverted; see Motowidlo & Beier, 2010). Agreement levels should be computed by comparing the ratings across judges and by comparing the ratings with a priori trait levels that the response options were designed to express.

In Figure 11.2, we present an example of a general domain knowledge SJT item that was taken from Motowidlo and colleagues (2006a). Contrary to contextualized SJTs (see Figure 11.1), the description of the situation is more generic and more widely applicable across many job situations and is specifically intended to serve as a framework for the measurement of a particular construct (in this case the personality trait agreeableness). Another difference is that the response options were specifically written to measure agreeableness. Whether the response options are indicative of high or low levels of agreeableness is mentioned in parentheses. People who rate those options that express high levels of the personality trait positively and those options that express low levels of the personality trait negatively are believed to be in high possession of the trait and have general domain knowledge about how to express this trait effectively in work situations.

## Overview of prior research

Understandably, so far there has been less research on general domain knowledge SJTs. In the next subsection, we follow the same structure as with context-specific SJTs. That is, we review the research evidence to date on reliability, criterion-related and incremental validity, construct-related validity, subgroup differences, applicant reactions, faking, retest and coaching effects.

*Reliability* The internal consistency reliability for domain-general SJT scores is not superior to context-specific SJT scores because it can be argued that domain-general SJTs also have a multidimensional nature. Although domain-general SJTs are designed to tap single personality traits, an expression of a trait like agreeableness, for example, could show some overlap with extraversion because extraversion could also be required to express agreeableness effectively in a particular situation (see also further below; Motowidlo et al., 2006a). Motowidlo and Beier (2010) reported internal consistency reliability estimates for a domain-general SJT tapping the personality dimensions of agreeableness, extraversion and conscientiousness ranging from 0.40 to 0.65, which is comparable to domain-specific SJTs. Motowidlo and colleagues (2009) further reported reliability estimates for their single-response SJT in the range of 0.21 to 0.55. Only one study so far has shown that, as with contextualized SJTs, alternative form reliability of domain-general SJT item scores tends to be higher ($r = 0.71$) than internal consistency reliability estimates (Motowidlo et al., 2006a).

*Criterion-related and incremental validity* As stated above, the theory of knowledge determinants underlying SJT performance builds on the premise that knowledge predicts actual behaviour in both simulated and actual workplace settings. Recent studies (Lievens & Patterson, 2011; Lievens & Sackett, 2012) provide empirical support for the conceptual link between knowledge and behaviour. In these studies, the relation between procedural knowledge as measured by an SJT and future job performance was mediated by either internship behaviour or, in the case of the second study, assessment centre performance (see also Crook et al., 2011; Kell, Motowidlo, Martin, Stotts & Moreno, 2014; Motowidlo, Martin & Crook, 2013). Importantly, domain-general SJTs also show correlations with job performance of a similar magnitude to traditional SJTs. Motowidlo and Beier (2010) reported correlations from 0.21 to 0.29 for ITP scoring keys with supervisory ratings of job performance. Recently, Motowidlo and colleagues (2013) found evidence indicating that knowledge about effective and ineffective behaviour predicted role-play simulation performance in handling service encounters and work effort performance over and above the personality traits of extraversion, conscientiousness, emotional stability and openness. Crook and colleagues (2011) found similar results: knowledge remained an important predictor of job performance after personality was accounted for.

*Construct-related validity* Research found that people's ratings on SJT response options that express high levels of the personality trait of conscientiousness or agreeableness show substantial correlations in the range of 0.40 to 0.50 with their corresponding personality trait scores as measured by self-reports (Motowidlo & Beier, 2010). For other personality dimensions, the evidence was less convincing. Extraversion, for example, correlated only 0.12–0.21 with the personality trait extraversion as measured by a self-report personality inventory. As suggested before, domain-general SJTs do not seem to solve the multidimensionality problems that characterize their contextualized counterparts. That is, behavioural content for one personality trait is potentially confounded by behavioural content expressing another personality trait (Motowidlo et al., 2006a). Consider the SJT example in Figure 11.2, and more specifically response option f), 'Address the group as a whole and state that it is important to keep comments constructive'. This option (especially the second half of the sentence) might represent high levels of agreeableness. However, one might interpret this response option (especially the first half of the sentence) as equally an expression of extraversion. So, even though general-domain knowledge SJTs can be specifically designed to measure a single trait (and a retranslation procedure with SMEs can be performed to verify this), the response options seem still saturated with more than one trait because of the well-known correlations among personality traits.

*Subgroup differences* No studies to date have tackled the question of whether there are subgroup differences in domain-general SJT scores and whether these are lower than those found with traditional SJTs. It can be expected that domain-general SJT scores reduce subgroup differences in comparison with contextualized SJTs because they mainly tap test-takers' ITPs and require little if any specific job experience. Domain-general SJTs are also presumed to be less cognitively saturated, which is the main driver behind subgroup differences in selection test scores. So, racial differences might be reduced in applicant pools that take a generic SJT. Gender differences, on the other hand, might increase when generic SJTs are used to specifically tap the personality traits conscientiousness and agreeableness, thereby giving women an advantage over men (Whetzel et al., 2008).

*Applicant reactions* Little is known about how applicants react to general domain knowledge SJTs. Given their generic nature, a key question for future research is to investigate if they are seen as sufficiently face-valid and job-related. We do not have empirical answers to these questions yet and therefore they remain to be answered in future research. Important moderators seem to be the sample (inexperienced vs. experienced applicants) and the SJT purpose (e.g., entry-level admission vs. advanced level testing). For instance, whereas inexperienced candidates may view generic SJT items as sufficiently face-valid, more experienced candidates may expect more contextualized information to apply to their fine-grained knowledge. Advanced level selection might also require more contextualization for the same reason.

*Faking, retest and coaching effects* Theoretically, general domain knowledge SJTs that tap ITPs are supposedly less prone to faking than more explicit measures of personality since they cannot be subjected to response distortion and social desirability as easily as self-report personality questionnaires (Motowidlo et al., 2006b). However, no published research evidence attesting to this argument has been found thus far. In addition, we are not aware of studies comparing the fakability of SJTs measuring ITPs to contextualized SJTs.

As indicated earlier, the theory of knowledge determinants underlying SJT performance states that SJTs measure procedural knowledge acquired when people are exposed to situations that provide opportunities for learning (Motowidlo et al., 2006b). So, the theory implies that performance on SJTs that capture this general domain knowledge might be trainable since people can develop their knowledge about the costs and benefits of expressing certain traits in particular (job-related) situations and this knowledge can then supplement or even override their inherent trait expressions. Some initial research has tested these assumptions. In particular, a recent study of medical students in India reported that their procedural knowledge scores reflecting ITP expressions increased throughout the medical curriculum (Ghosh, Motowidlo & Nath, 2014).

## General Domain Knowledge SJTs: Implications and Trends

The conceptualization of SJTs as measures of relatively context-independent knowledge has fundamental implications for SJT design. If SJTs aim to tap into general domain knowledge, it seems to make less sense to invest in elaborate, contextualized situation descriptions. Instead, this perspective conceptually guides research efforts to streamline SJTs. This can be done in at least two ways. One approach is to make use of single-response SJTs in which test-takers are asked to rate the effectiveness of a single critical action (Crook et al., 2011). As described above, traditional contextualized SJT items usually have multiple-response options. Test developers have to gather a large number of response options from SMEs in the test construction phase. Next, response options have to be investigated and checked for redundancy and SME agreement, before ending up with a pool of suitable response options for the final SJT. Single-response SJTs are proposed to reduce this laborious and time-intensive process because the edited critical incidents (i.e., retaining only the situation description and a single critical action) can directly serve as the response options, thereby rendering the need for generating large amounts of response options superfluous. SMEs also simply rate the effectiveness of edited critical incidents. When applicants complete the SJT, they have to provide an effectiveness rating for each item, which is compared to the one generated by the SME for scoring purposes. Thus, each item of a single-response SJT consists of a couple of sentences describing one critical brief incident, with candidates being asked to rate the effectiveness of this incident. Crook and colleagues (2011) created such single-response SJTs (see also Motowidlo et al., 2009; Motowidlo et al., 2013). In two studies, Crook et al. (2011) found single-response SJTs to be significantly correlated with performance ($r$ = 0.22–0.33), and showed that job knowledge as measured by one of their SJTs showed 4% incremental variance on SJT scores above personality. These preliminary findings are in line with McDaniel and colleagues' (2007) meta-analytic evidence of traditional SJTs, suggesting that single-response SJTs do not appear to pay a 'predictive power reduction price' for their streamlined development.

A similar approach would be to eliminate the situation stem altogether and ask test-takers to rate the effectiveness of several courses of action from a multiple-choice response option set (Kell, Martin & Motowidlo, 2011) Kell et al. (2014) devised such a test consisting of 40 brief descriptions of courses of action – in this case physicians interacting with patients. An example of one such (effective) description is: 'When a 10 year old with a broken arm needed surgery, the anesthetist introduced herself to the parents and then knelt down to the child's eye level to introduce herself to the child'. The statements were developed from critical incidents. Test-takers have to score each item's effectiveness. Thus, their test is similar in format to single-response SJTs (with the exception that only the actions were retained and the situations were dropped from the items) and was designed to measure prosocial knowledge (i.e., helping behaviour). Prosocial knowledge as measured with this instrument correlated 0.20 with clinical skill on a standardized patient examination (SPE). Furthermore, prosocial knowledge scores were positively associated with students' clinical performance scores from their primary care rotations ($r$ = 0.22), but non-significantly correlated to students' clinical performance scores in the specialties ($r$ = –0.04). So, this study suggests that general-domain knowledge seems to be more important in the early phases of one's career before specialization takes place, and declines in importance in the later phases when specialized skills become more and more essential.

## Suggestions for Future Research and Recommendations for Practice

After outlining two SJT perspectives (context-dependent vs. general domain knowledge), we end this chapter by highlighting some important avenues for future research. In the preceding sections, we briefly touched on some of those future research directions.

A first vital issue is to gain a better understanding of the circumstances in which each perspective produces the best criterion-related and construct-related validity evidence. For example, when designing an SJT for entry-level admission purposes, evidence accumulated throughout this chapter is unsupportive of contextualizing such an SJT and instead supports streamlining the SJT, thereby making it more context-independent (e.g., Kell et al., 2014). Interestingly, development costs are reduced while at the same time the test's criterion-related and construct-related validity are not jeopardized. As a potential disadvantage, however, applicants might perceive the generic SJT to be less job-related because the relation to the job becomes somewhat less obvious (as manifested in the more generic wording of the item stems and response options). Similarly, at this time we do not know how contextualized and domain-general SJTs compare to one another in terms of fakability, subgroup differences and coachability. In such comparative evaluations, it is important to take the method–construct distinction into account (Arthur & Villado, 2008). That is, when tests are compared on their content, the test format should be kept constant. By directly contrasting contextualized with domain-general SJTs, it becomes possible to provide HR practice with the empirical evidence it needs to confirm the legitimacy of these issues. Krumm and colleagues (2015) carried out an example of such a study. They distinguished between two conditions: one in which a traditional SJT was used and another condition in which the situation description was removed from the items of the same SJT. So, respondents received only the item options in that condition. These conditions were implemented across three SJTs: a teamwork SJT, an integrity SJT and an aviation SJT. The results showed that the provision of context had less impact than expected. That is, it did not matter for about 50–70% of the items whether situation descriptions were included in terms of the number of correct solutions per item. In addition, respondents' expertise level, item length, item difficulty and response instruction did not moderate the results.

A second area of research deals with examining the effectiveness of the different SJTs for specific practical purposes. As a backdrop to this, we recommend using domain-general SJT items for entry-level selection. Conversely, context-specific SJTs seem particularly useful when applicants have already acquired the requisite fine-grained procedural (general) and declarative (job-specific) knowledge. Contextualized SJT items can then home in on such context-dependent knowledge. These items are particularly useful for advanced-level selection and certification applications. When selecting for specialized functions, declarative knowledge is an essential component in addition to procedural knowledge for effective job performance. Initial research is supportive of these recommendations because in advanced-level selection, administering a contextualized SJT was found to capture both procedural and declarative knowledge (Lievens & Patterson, 2011). Future studies should focus on further elucidating the additional value of increasing the contextualization of SJTs in advanced-level selection as compared to domain-general SJTs.

Training applications represents another SJT purpose that is relevant to our distinction and in urgent need of research. For training purposes, we also recommend using contextualized SJTs. SJTs might be specifically adapted to use as tools in training needs analysis (assessment of pre-training knowledge), as actual training content materials or as a training outcome assessment instrument. In particular, contextualized SJTs might be useful as training content materials in scenario-based training in which scripted work situations allow trainees to practice critical job-related skills in a safe environment (Fritzsche, Stagl, Salas & Burke, 2006). So far, virtually no research is available on the efficacy of using SJTs in training. Therefore, we need studies that explore to what extent increasing the response and/or stimulus fidelity of SJTs improves the training's effectiveness.

Fourth, future research might benefit from making a clearer distinction between these two SJT types. Many existing SJTs contain both generic and contextualized items

(Krumm et al., 2015). This might impact construct measurement. In particular, keeping the contextualization of SJT items at the same level (as required by the criterion specificity to be predicted) might lead to a better measurement model underlying SJTs as some of the item heterogeneity that has been posited to lead to poor factor analytical results in SJTs is removed. Generally, we believe that SJT research should not receive a 'free pass' on the construct measurement issue and should continue to undertake efforts to improve construct measurement in SJTs.

Efforts on a clearer distinction between these two SJT types might also address when and how test-takers make use of the context provided. That is, we should also be concerned with the underlying thought processes when solving SJTs. Leeds (2012) suggests that solving an SJT is a two-step process in which test-takers first scrutinize response alternatives in an absolute ('How effective is this option?' 'Does it make sense?') as well as in a relative sense ('Is this option better than that one?'). In a second process, test-takers take the contextual information as presented in the situation description into account. So, one may assume that even contextualized SJTs are only 'used' for context-specific judgements if test-takers' primary perusal of response options is inconclusive as to how to respond. Interestingly, Rockstuhl, Ang, Ng, Lievens and van Dyne (2015) revealed that the judgements made by test-takers on the basis of the situation descriptions (i.e., their construal of the situation) were equally or even more predictive of job-related criteria in an international context as compared with the judgements made on the basis of response alternatives alone. Thus, how test-takers construe and use the context provided could also be an important part of the information captured with SJTs. An avenue for future research may be to comparatively examine the cognitive underpinnings described by Leeds (2012) and Rockstuhl and colleagues (2015) (e.g., through eye-tracking or verbal protocol analysis) and also to assess their relevance in contextualized and generic SJT items.

A final interesting aspect of context-independent SJTs that deserves more research deals with their claim that they can be used cross-culturally. This assumption is based on the notion that such SJTs were designed to measure general procedural knowledge of the costs and benefits of engaging in specific trait-relevant behaviour. Conversely, contextualized SJTs are more dependent on the context and culture for which they were developed, and therefore cross-cultural transportability might be a problem (Lievens, 2006). Like cognitive ability and personality tests, general domain knowledge SJTs are developed to have generalizability across a wide variety of situations. Therefore, they could potentially be implemented more easily across different cultures. That said, we also caution that ITPs might be valued differently across cultures. For example, individualistic cultures might value expressions of extraversion in a specific situation, whereas collectivistic cultures might value these expressions less in that same situation and instead value expressions of other traits such as agreeableness more. Accordingly, empirical evidence is needed to determine the extent to which domain-general SJTs can be successfully implemented across different cultures.

## Conclusion

This chapter delineates two perspectives about the determinants of SJT performance: the contextualized perspective views SJTs as measures of job-specific knowledge, whereas the other perspective views SJTs as measures of general domain knowledge. Many current SJTs are situated somewhere between the two. Both perspectives are useful but have different SJT design implications. One perspective suggests further investing in more realistic stimulus and response formats. Conversely, the other perspective suggests streamlining SJTs. An important practical implication of the first perspective is the promise of improved predictive power involved in more realistic SJTs, while the second perspective posits that criterion- and construct-related validity would not suffer and indeed could benefit from designing more generic SJTs allowing broader predictions. This might especially hold for entry-level selection purposes because contextualization appears to be of higher importance for advanced-level selection. In the future, it seems beneficial that a clearer demarcation is used between these two perspectives. We also provide recommendations for practice and a research agenda for more comparative research between these two SJT perspectives in terms of key selection variables.

## References

Arthur, W. J., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgement 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.

Arthur, W. J., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*, 435–442. doi: 10.1037/0021-9010.93.2.435.

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black–White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology*, *66*, 91–126. doi:10.1111/Peps.12007

Bruk-Lee, V., Drew, E. N., & Hawkes, B. (2014). Candidate reactions to simulations and media-rich assessments in personnel selection. *In* M. S. Fetzer & K. Tuzinski (Eds.), *Simulations for Personnel Selection* (pp. 43–60). New York: Springer.

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgement measures: Interactionist psychology operationalized. *In* N. D. Christiansen & R. P. Tett (Eds.), *Handbook of Personality at Work* (pp. 439–456). New York: Routledge.

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

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgement tests used in high-stakes situations. *International Journal of Selection and Assessment*, *20*, 333–346. doi: 10.1111/j.1468-2389.2012.00604.x.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgement 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 judgement 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.

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternate predictors and an example. *Personnel Psychology*, *51*, 193–208. doi: 10.1111/j.1744-6570.1998.tb00722.x.

Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance using single-response situational judgement tests. *International Journal of Selection and Assessment*, *19*, 363–373. doi: 10.1111/j.1468-2389.2011.00565.x.

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgement tests in the college admission process. *International Journal of Selection and Assessment*, *14*, 142–155. doi: 10.1111/j.1468-2389.2006.00340.x.

De Soete, B., Lievens, F., Oostrom, J. K., & Westerveld, L. (2013). Alternative predictors for dealing with the diversity-validity dilemma in personnel selection: The constructed response multimedia test. *International Journal of Selection and Assessment*, *21*, 239–250. doi: 10.1111/ijsa.12034.

Dunlop, P. D., Morrison, D. L., & Cordery, J. L. (2011). Investigating retesting effects in a personnel selection context. *International Journal of Selection and Assessment*, *19*, 217–221. doi: 10.1111/j.1468-2389.2011.00549.x.

Fetzer, M. S., & Tuzinski, K. (2014). *Simulations for Personnel Selection*. New York: Springer.

Fetzer, M. S., Tuzinski, K., & Freeman, M. (2010). 3D animation, motion capture, and SJTs: I-O is finally catching up with it. Paper presented at the 25th Annual Conference of Industrial and Organizational Psychology, April. Atlanta, GA.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*, 327–358. doi: 10.1037/h0061470.

Fritzsche, B. A., Stagl, K. C., Salas, E., & Burke, C. S. (2006). Enhancing the design, delivery, and evaluation of scenario-based training: Can situational judgement tests contribute? *In* J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement, and Application* (pp. 301–318). Mahwah, NJ: Lawrence Erlbaum.

Ghosh, K., Motowidlo, S. J., & Nath, S. (2014). Effects of prosocial and technical knowledge on students' clinical performance. Paper presented at the 29th Annual Conference of the Society for Industrial and Organizational Psychology, May, Honolulu, Hawaii.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, *57*, 639–683. doi: 10.1111/j.1744-6570.2004.00003.x.

Hooper, A. C., Cullen, M. C., & Sackett, P. R. (2006). Operational threats to the use of SJTs: Faking, coaching, and retesting issues. *In* J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement, and Application* (pp. 205–232). Mahwah, NJ: Lawrence Erlbaum.

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of situational judgement items. *European Journal of Psychological Assessment*, *22*, 168–176. doi: 10.1027/1015-5759.22.3.168.

Kasten, N., & Freund, P. A. (2015). A meta-analytical multilevel reliability generalization of situational judgement tests (SJTs). *European Journal of Psychological Assessment*. doi: 10.1027/1015-5759/a000250.

Kell, H. J., Martin, M. P., & Motowidlo, S. J. (2011). Medical students' knowledge about medical professionalism predicts their professionalism performance. Poster presented at the 26th Annual meeting of the Society for Industrial and Organizational Psychology, April, Chicago, IL.

Kell, H. J., Motowidlo, S. J., Martin, M. P., Stotts, A. L., & Moreno, C. A. (2014). Testing for independent effects of prosocial knowledge and technical knowledge on skill and performance. *Human Performance*, *27*, 311–327.

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How 'situational' is judgement in situational judgement tests? *Journal of Applied Psychology*, *100*, 399–416. doi: 10.1037/a0037674.

Leeds, J. P. (2012). The theory of cognitive acuity: Extending psychophysics to the measurement of situational judgement. *Journal of Neuroscience, Psychology, and Economics*, *5*, 166–181. doi: 10.1037/a0027294.

Lievens, F. (2006). International situational judgement tests. *In* J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement, and Application* (pp. 279–300). Mahwah, NJ: Lawrence Erlbaum.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*, 981–1007. doi: 10.1111/j.1744-6570.2005.00713.x.

Lievens, F., Buyse, T., Sackett, P. R., & Connelly, B. S. (2012). The effects of coaching on situational judgement tests in high-stakes selection. *International Journal of Selection and Assessment*, *20*, 272–282. doi: 10.1111/j.1468-2389.2012.00599.x.

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management*, *41*(6), 1604–1627. doi: 10.1177/0149206312463941.

Lievens, F., & De Soete, B. (2012). Simulations. *In* N. Schmitt (Ed.), *The Oxford Handbook of Personnel Assessment and Selection* (pp. 383–410). Oxford: Oxford University Press.

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology*, *96*, 927–940. doi: 10.1037/A0023496

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

Lievens, F., & Sackett, P. R. (2007). Situational judgement tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology*, *92*, 1043–1055. doi: 10.1037/0021-9010.92.4.1043.

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgement tests for predicting academic success and job performance. *Journal of Applied Psychology*, *97*, 460–468. doi: 10.1037/A0025741.

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

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgement 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 judgement tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*, 103–113. doi: 10.1111/1468-2389.00167.

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Week, J. A. (2011). Toward an understanding of situational judgement item validity and group differences. *Journal of Applied Psychology*, *96*, 327–336. doi: 10.1037/a0021983.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system-theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268. doi: 10.1037/0033-295x.102.2.246.

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgement test. *Journal of Applied Psychology*, *95*, 321–333. doi: 10.1037/A0017975.

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

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*, 640–647. doi: 10.1037/0021-9010.75.6.640.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). Implicit policies about relations between personality traits and behavioural effectiveness in situational judgement items. *Journal of Applied Psychology*, *91*, 749–761. doi: 10.1037/0021-9010.91.4.749.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). A theoretical basis for situational judgement tests. *In* J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement, and Application* (pp. 57–82). Mahwah, NJ: Lawrence Erlbaum.

Motowidlo, S. J., Martin, M. P., & Crook, A. E. (2013). Relations between personality, knowledge, and behaviour in professional service encounters. *Journal of Applied Social Psychology*, *43*, 1851–1861. doi: 10.1111/jasp.12137.

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

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology*, *19*, 532–550. doi: 10.1080/13594320903000005.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*, 343–355. doi: 10.1037//1082-989x.5.3.343.

Patterson, F., Ferguson, E., Norfolk, T., & Lane, P. (2005). A new selection system to recruit general practice registrars: Preliminary findings from a validation study. *British Medical Journal*, *330*, 711–714. doi: 10.1136/bmj.330.7493.711.

Peeters, H., & Lievens, F. (2005). Situational judgement tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*, 70–89. doi: 10.1177/0013164404268672.

Potosky, D. (2008). A conceptual framework for the role of the administration medium in the personnel assessment process. *Academy of Management Review*, *33*, 629–648. doi: 10.5465/AMR.2008.32465704.

Richman-Hirsch, 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.

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgement tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology*, *100*, 464–480. doi: 10.1037/a0038098.

Roth, P. L., Bobko, P., & Buster, M. A. (2013). Situational judgement tests: The influence and importance of applicant status and targeted constructs on estimates of Black–White subgroup differences. *Journal of Occupational and Organizational Psychology*, *86*, 394–409. doi: 10.1111/Joop.12013.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274. doi: 10.1037//0033-2909.124.2.262.

Schmitt, N., & Chan, D. (2006). Situational judgement tests: Method or construct? *In* J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgement Tests: Theory, Measurement, and Application* (pp. 135–155). Mahwah, NJ: Lawrence Erlbaum.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the behavioural consistency approach: Selection test development based on a content-oriented strategy. *Personnel Psychology*, *39*, 91– 108. doi: 10.1111/j.1744-6570.1986.tb00576.x.

Stemig, M., Sackett, P. R., & Lievens, F. (2015). Effects of organizationally-endorsed coaching on performance and validity of situational judgement tests. *International Journal of Selection of Assessment*, *23*, 175–182.

Thornton, G. C., & Rupp, D. E. (2006). *Assessment Centres in Human Resource Management: Strategies for Prediction, Diagnosis, and Development.* Mahwah, NJ: Lawrence Erlbaum.

Wagner, R. K., & Sternberg, R. J. (1991). *Tacit Knowledge Inventory for Managers (TKIM).* New York: Psychological Corporation.

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behaviour*, *2*, 295–322. doi: 10.1146/annurev-orgpsych-032414-111304.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, *52*, 679–700. doi: 10.1111/j.1744-6570.1999.tb00176.x.

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

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples and criteria. *Journal of Applied Psychology*, *52*, 372–376. doi: 10.1037/H0026244.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgement tests: An overview of current research. *Human Resource Management Review*, *19*, 188–202. doi: 10.1016/j.hrmr.2009.03.007.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgement test performance: A meta-analysis. *Human Performance*, *21*, 291–309. doi: 10.1080/08959280802137820.