



Situational judgment tests: a review of recent research

Filip Lievens, Helga Peeters and Eveline Schollaert
Ghent University, Ghent, Belgium

Received 4 December 2006
Revised February 2007
Accepted 24 May 2007

Abstract

Purpose – The purpose of this paper is to give an empirically-based review of the strengths and weaknesses of situational judgment tests (SJTs).

Design/methodology/approach – The features, history, and development of SJTs are discussed. Next, a computerized search (from 1990-2007) is conducted to retrieve empirical studies related to SJTs. The effectiveness of SJTs is discussed in terms of reliability, criterion-related validity, incremental validity, construct-related validity, utility, adverse impact, applicant perceptions, fakability, and susceptibility to practice and coaching effects.

Findings – Strengths of SJTs are that they show criterion-related validity and incremental validity above cognitive ability and personality tests. SJTs have also less adverse impact towards minorities (especially if the cognitive loading of the SJT is low). Furthermore, applicant reactions towards SJTs are positive and SJTs enable to test large applicant groups at once (through the Internet). In terms of weaknesses, SJTs might be prone to faking, practice, and coaching effects. There is also debate about what constructs are measured by SJTs.

Research limitations/implications – Five avenues for future research are discussed: construct-related validity of SJTs, utility of SJTs *vis-à-vis* other predictors, impact of SJT features on validity and adverse impact, examination of alternative stimulus and response formats, and cross-cultural transportability of SJTs.

Practical implications – Practitioners receive evidence-based information about the features, development, and strengths and weaknesses of SJTs.

Originality/value – Apart from the USA, SJTs have not made strong inroads in selection practice in Europe and other parts of the world. This evidence-based paper might highlight the value of SJTs.

Keywords Selection, Interpersonal skills, Human resourcing, Aptitude tests, Competences

Paper type General review

In recent years, many customers for whom a recruitment office was working have requested more detailed information on the interpersonal competencies of candidates in the first selection stage. However, they were sceptical about using self-report personality inventories. The recruitment office was looking for an efficient way of measuring interpersonal competencies in preliminary selection stages.

The Armed Forces were facing a high attrition rate among newly hired officers. Apparently, the officers hired had a too romanticized image of the Armed Forces as an employer. The Armed Forces wanted to include a realistic job preview on the recruitment web site. So, a test was put on the web site wherein potential applicants for officer jobs were given short military leadership scenarios (including pictures) and were asked what to do. Feedback on the correct answer was provided.

In many countries, cognitively-oriented predictors are typically used in admission exams. Although these cognitive tests are valid for predicting academic performance, they also exhibit large adverse impact. One is seeking standardized tests that broaden the competencies assessed and that can be administered to large groups of people.



These three situations are only some examples which illustrate why situational judgment tests (SJTs) have become increasingly popular in personnel selection in the USA (Ployhart, 2006). In a nutshell, as contextualized measurement methods SJTs are attractive to practitioners because they enable to measure mostly interpersonally-oriented skills among a large group of people in preliminary selection stages. Similarly, in recent years empirical research with regard to SJTs has been flourishing (Weekley and Ployhart, 2006). Despite these developments in practice and research, SJTs have not made strong inroads in selection practice in Europe and other parts of the world. For example, the most recent worldwide survey of the usage of selection practices (Ryan *et al.*, 1999) shows that in virtually all countries SJTs were substantially less used than more common selection procedures such as cognitive ability tests and personality inventories. Only in two of the 20 countries (the USA and Hong Kong) this was not the case. In addition, it is striking that important handbooks on personnel selection (e.g. Cook, 2003) do not include a chapter on SJTs. One of the reasons might be that practitioners are relatively unfamiliar with SJTs as selection procedures. Another reason might be that practitioners are familiar with SJTs but that they are sceptical about their effectiveness.

As it seems that SJTs are neither widely known nor used the aim of this paper is two-fold. First, we present the “nuts and bolts” of SJTs. This includes questions such as: What are SJTs? How do they differ from other selection procedures like assessment centre (AC) exercises? How can they be developed? Second, we discuss the empirical evidence behind SJTs. In other words, what are the empirically-based strengths and weaknesses of SJTs?

Situational judgment tests

Definition, main features, and history of SJTs

SJTs are measurement methods that present applicants with job-related situations and possible responses to these situations. Applicants have to indicate which response alternative they would choose. SJTs have a long history (McDaniel *et al.*, 2001). Similar to ACs, the origin of SJTs goes back to civil service and military examinations. There are also traces of SJT use during World War II. The modern version of the SJT was “invented” by Motowidlo *et al.* (1990). These “modern” SJTs share the following characteristics (McDaniel and Nguyen, 2001; Weekley and Ployhart, 2006). First, SJTs present applicants with job-related situations. The notion that the situations presented are related to the job (or a family of jobs) increases the job-relatedness of SJT items. However, SJTs may vary in terms of the fidelity with which they present the stimuli (i.e. the situations). The “fidelity of the task stimulus” refers to the extent to which the format of the task stimulus is consistent with how the situation is encountered in the workplace. Along these lines, a distinction is made between written SJTs on the one hand, and video-based or multimedia SJTs on the other hand. Regarding the former, an SJT takes the form of a written test as the scenarios are presented in a written format and applicants are asked to indicate the appropriate response alternative. Thus, written SJTs seem to have low stimulus fidelity. Conversely, a video-based test consists of a number of video scenarios. Each scenario describes a person handling a typical job-related situation. At a critical “moment of truth”, the scenario ends and the applicant is asked to choose among several courses of action (Dalessio, 1994; Smiderle *et al.*, 1994; Weekley and Jones, 1997). Questions and response options are presented

visually and supported by narration. Multimedia SJTs are basically the same as video-based SJTs. The only exception is that CD or DVD serve as the medium (instead of VCR, McHenry and Schmitt, 1994). Video-based and multimedia SJTs allow the item context to be richly portrayed, increasing their stimulus fidelity and – as will be indicated below – their validity (Funke and Schuler, 1998).

Second, SJTs have a multiple-choice item format. This means applicants have to choose an alternative from a list of response options. Again, the response alternatives can be presented in either a written (low response fidelity) or a video-based (medium response fidelity) format. In any case, applicants are not asked to show or even report on actual behaviour (high fidelity). This feature discriminates SJTs from high-fidelity simulations such as AC exercises, which provide applicants with the opportunity to respond in a manner mimicking actual job behaviour observed and evaluated by trained assessors. Note that an SJT virtually resembles a situational interview when situations with an open-ended response format are presented orally to candidates.

A couple of other SJT response modalities deserve attention. Sometimes the applicant's response to a situation determines the next situation that is presented. So, applicants are confronted with possible consequences of their choices. This modality implies that all applicants are not asked to respond to the same items. These SJTs are called "branched", "nested", or "interactive" SJTs (Olson-Buchanan *et al.*, 1998). The technological possibility of developing interactive SJTs is inherent in multimedia SJTs which present different video fragments to an applicant, based on the applicant's response to earlier video fragments. This allows the SJT to simulate the dynamics of interaction, while maintaining some levels of standardization (see below, for research on the validity of video-based SJTs). Apart from branching, another response modality is that the response instructions of SJTs can have either a knowledge format ("What is the best answer?") or a behavioural tendency format ("What are you most likely to do?"). As noted by McDaniel *et al.* (2007), SJTs with knowledge response instructions measure maximal performance. Similar to cognitive ability or job knowledge tests, in SJTs with knowledge response instructions candidates are motivated to show whether they know what the most effective answer is. Conversely, behavioural tendency instructions measure typical performance because they require candidates to report how they typically behave, which is similar to personality inventories.

A final characteristic of SJTs is that the scoring key is developed a priori. This means that there are no assessors or raters who evaluate candidates' behaviour. Along these lines, SJTs share many parallels with biodata inventories. Similar to biodata scales, SJTs are often scored on the basis of experts or empirical grounds (Bergman *et al.*, 2006). We discuss this issue at length in the following section.

In conclusion, these features of SJTs clarify the parallels and differences between SJTs and better-known predictors such as AC exercises (for a summary, see Table I). Similarities between SJTs and ACs include that they both build on the behavioural consistency and psychological fidelity principles. In addition, they are both methods which might capture a multitude of constructs. Differences include that SJTs might be administered to large groups and over the Internet, whereas ACs are typically used in smaller samples in a specific location. As noted above, AC exercises are high-fidelity simulations because trained assessors observe and rate actual ongoing candidate behaviour. Conversely, SJTs are low-fidelity simulations as candidates have to pick the "correct" answer from a limited set of predetermined response options. Accordingly,

	SJTs	Assessment centres
Type of simulation	Low-fidelity simulation for measuring a variety of constructs	High-fidelity simulation for measuring a variety of constructs
Stimulus	Contextualized and job-related Standardized written or video-based situations	Contextualised and job-related Standardized content and instructions
Response	Written response Self-report Multiple-choice format	Unexpected behaviour of other candidates and/or role-players Behavioural response Other-report (i.e. assessors) Open ended format
Scoring	A priori determined (expert-based or empirically-derived)	“Live” (or video) observation and rating by trained assessors
Use	Select-out Large groups (internet)	Select-in Small groups

Table I.
Comparison between
SJTs and assessment
centre exercises

standardization is ensured because everybody receives the same situations (with the exception of branched SJTs) and the scoring key is determined *a priori*.

Development of SJTs

As described by Motowidlo *et al.* (1990), the development of SJTs typically follows three stages. The development starts with a job analysis wherein critical incidents of work situations are collected from subject matter experts (e.g. incumbents, supervisors, customers) or in some cases from archival sources. SJT developers often aim to gather incidents that deal with specific content domains or constructs related to the job (Bergman *et al.*, 2006). However, as noted below, this does not mean that these content domains are retrieved in factor analyses of SJT items. Upon gathering critical incidents, the test developer then groups the incidents, selects representative scenarios, and edits the incidents into stems of similar length and format.

In a second step, a different group of subject matter experts or inexperienced employees is asked to generate one or more responses to each situation. Subject matter experts are useful because they should be able to identify the best responses and to generate some less optimal responses. Inexperienced employees are useful because they offer responses with a wide range of effectiveness. After gathering response alternatives, the SJT developer edits all of the response alternatives so that the responses of an item span a range of effectiveness.

In a final step, the scoring key is developed. Although there are various scoring methods for multiple-choice SJTs (Bergman *et al.*, 2006), “rational” and “empirical” scoring approaches are typically distinguished. When SJT items are rationally scored, experts (excellent employees) are asked to make judgments concerning the effectiveness of the responses, or they identify the best and the worst options. Options identified as “best” are scored as correct and options identified as “worst” are scored as incorrect. When SJT items are empirically scored, they are administered to a large pilot sample. Next, items (or response alternatives) are selected and/or weighted according to evidence that the items (or response alternatives) differentiate between persons who score at different levels on a criterion variable (e.g. job performance). Response options that are relatively often selected by individuals who perform highly

on the criterion are scored as correct. Options selected relatively often by low performing individuals are scored as incorrect. In some cases, a hybrid scoring scheme is used which combines rational and empirical scoring.

Strengths and weaknesses of SJTs: an evidence-based overview

In this section, we discuss potential benefits and drawbacks of SJTs on the basis of a review of extant empirical research on SJTs. To this end, we searched for empirical studies related to SJTs using a number of computerized databases (e.g. Web of Science). In terms of publication year range, we searched for studies from 1990 (publication of Motowidlo *et al.*) until 2007. We also scrutinized reference lists from obtained studies to find other published and unpublished studies. This resulted in a list of 52 articles.

In the remainder, the effectiveness of SJTs is reviewed on the basis of the following well-known “hard” and “soft” psychometric criteria: reliability, criterion-related validity, incremental validity, construct-related validity, utility, adverse impact, applicant perceptions, fakability, and susceptibility to practice and coaching effects. Table II summarizes the results of our evidence-based review of SJTs.

Reliability

Most prior research on the reliability of SJTs has examined the internal consistency reliability of SJTs. The meta-analysis of McDaniel *et al.* (2001) summarized these studies and found that the internal consistency coefficients varied between 0.43 and 0.94. Research identified various factors that moderate this variability in internal consistency reliability. Logically, the length of the SJT played a role, with longer SJTs showing higher internal consistency. In addition, Ployhart and Ehrhart (2003) found that the type of response instructions influenced the internal consistency. Asking candidates “to rate the effectiveness of each response” led to the highest internal consistency (0.73). Response instructions wherein candidates had to choose two response alternatives (“Pick the best and worst response”) had somewhat lower internal consistency (0.60), whereas response instructions wherein candidates had to choose only one response (e.g. “What is the most effective response?”) had the lowest internal consistency (0.24).

The findings that SJTs might have low internal consistencies and that they are multidimensional are also evidenced by factor analyzing SJT items. Such factor analytic SJT research typically reveals a plethora of factors that are difficult to interpret (Schmitt and Chan, 2006). This is not surprising as SJTs are measurement methods which assess a variety of work-related knowledge, skills, and abilities (KSAs) (McDaniel *et al.*, 2001; McDaniel and Whetzel, 2005; Weekley and Jones, 1999). For instance, SJTs were recently developed to capture domains as diverse as teamwork knowledge (McClough and Rogelberg, 2003; Morgeson *et al.*, 2005; Stevens and Campion, 1999), aviation pilot judgment (Hunter, 2003), employee integrity (Becker, 2005), call-centre performance (Konradt *et al.*, 2003), or academic performance (Oswald *et al.*, 2004).

The multidimensional nature of SJTs calls into question whether internal consistency is a good reliability measure for SJTs as internal consistency is only an adequate measure of reliability for unidimensional tests (McDaniel and Nguyen, 2001; Motowidlo *et al.*, 1990). It has been suggested that test-retest reliability is a better

Issue	Research findings	Major sources
Reliability	The internal consistency of SJTs varies from 0.43 to 0.94 SJTs multidimensionality, length, and response instructions affect the internal consistency	McDaniel <i>et al.</i> (2001) Ployhart <i>et al.</i> (2004)
Criterion-related validity	SJTs factor analyses show that SJTs are multidimensional Test-retest reliability of SJTs is adequate SJTs have a mean corrected correlation of 0.34 with job performance	Schmitt and Chan (2006); McDaniel and Whetzel (2005) Ployhart <i>et al.</i> (2004) McDaniel <i>et al.</i> (2001)
Incremental validity	Validity has been established in a variety of settings Video-based SJTs are more valid than written SJTs SJTs explain additional variance over cognitive ability, personality, job knowledge, and experience for predicting task performance SJTs explain additional variance over cognitive ability for predicting contextual performance	Liviens and Sackett (2006a) Chan and Schmitt (2002); Clevenger <i>et al.</i> (2001); Weekley and Jones (1997, 1999) McDaniel <i>et al.</i> (2007); O'Connell <i>et al.</i> (2007)
Construct-related validity	SJTs are measurement methods for assessing a variety of constructs	Schmitt and Chan (2006); McDaniel and Whetzel (2005)
Utility	Correlation with personality and cognitive ability depends on the constructs measured and on the response instructions used No research available	McDaniel <i>et al.</i> (2007)
Adverse impact	Whites score 0.38 <i>SD</i> better than Blacks Females score 0.10 <i>SD</i> better than males	Nguyen <i>et al.</i> (2005a) Nguyen <i>et al.</i> (2005a) Nguyen <i>et al.</i> (2005b)
Applicant perceptions	SJTs with behavioural tendency instructions have less adverse impact than SJTs with knowledge instructions Video-based SJTs have less adverse impact than written SJTs Video-based (multimedia) SJTs have more positive perceptions than written SJTs	Chan and Schmitt (1997) Chan and Schmitt (1997); Kanning <i>et al.</i> (2006); Richman-Hirsch <i>et al.</i> (2000) Hooper <i>et al.</i> (2006)
Fakability	Faking might improve SJT scores from 0.08 to 0.89 <i>SD</i> Faking has less impact on SJTs than on personality inventories	Hooper <i>et al.</i> (2006)
Susceptibility to coaching/practice	Item transparency, cognitive loading of the SJT, type of response instructions, and type of design moderate SJT fakability SJTs are susceptible to coaching and practice effects	Cullen <i>et al.</i> (2006)

Table II.
Overview of research findings with respect to SJTs

measure for assessing the reliability of SJTs. Several studies have scrutinized the test-retest reliability of SJTs. For instance, Ployhart *et al.* (2004) reported a test-retest reliability of 0.84. Bruce and Learner (1958) and Richardson, Bellows, Henry & Co. (1981) found test-retest reliabilities that ranged from 0.77 to 0.89 for the “Supervisory Practices Test” and for the “Supervisory Profile Record”. In short, these early and recent studies show that the test-retest reliability of SJTs (with sufficient length) is satisfactory.

Criterion-related validity

A key question in selection practice is whether a selection procedure is able to predict job-related criteria. Various studies have examined whether SJTs are good predictors of job performance. McDaniel *et al.* (2001) conducted the first meta-analysis of the criterion-related validities of SJTs (across 95 studies) in employment settings. The corrected correlation between SJTs and job performance was 0.34 (uncorrected correlation = 0.26). Furthermore, the substantial variability in criterion-related validity coefficients across studies suggested the presence of moderators. A key moderator of the validity of SJTs concerned whether a job analysis was used to develop the SJT; SJTs based on a job analysis evidenced higher validities than those not based on a job analysis (0.38 versus 0.29). Apart from the good validity of SJTs in employment settings, recent research has also shown that SJTs can be valid predictors in educational contexts (as part of admission exam testing, Lievens *et al.*, 2005a, b; Oswald *et al.*, 2004).

There are three general assumptions underlying why SJTs predict job performance (Motowidlo *et al.*, 1990). The first explanation indicates that the best predictor of future behaviour is past behaviour (i.e. the behavioural consistency principle). Thus, the assumption is that candidate performance on the selection instrument (SJT) will be consistent and therefore will be predictive of candidate performance on the job. A second assumption is that SJTs measure applicants’ intentions and goals (“goal-setting theory”). The final explanation is that SJTs measure constructs that have been shown to be pervasive, robust or useful predictors of job performance, like procedural knowledge, practical intelligence, general cognitive ability or personality traits (see below). At present, there is still considerable speculation about the exact mechanisms through which SJTs are related to job performance.

The issue of whether written SJTs or video-based (multimedia) SJTs are better predictors of job performance is an interesting one. On the one hand, video-based and multimedia SJTs might have higher fidelity because the presented information is richer and more detailed, which in turn might lead to a better match with the criterion behaviour as presented on the job. This might result in higher criterion-related validity. However, on the other hand, as cognitive ability is an important predictor of job performance, video-based and multimedia SJTs might be less valid because they are less cognitively loaded (i.e. lower reading component). Furthermore, the video-based format might insert irrelevant contextual information and bring more error into SJTs, resulting in lower validity. Lievens and Sackett (2006a) tried to test these assertions. They demonstrated that changing an existing video-based SJT to a written one (keeping content constant) significantly reduced the criterion-related validity of the test. In addition, the written version had a significantly higher correlation with cognitive ability.

Incremental validity

Apart from each predictor's validity, it is both theoretically and practically pivotal to examine the predictive validity of SJTs over other more established predictors. Typically referred to as incremental validity, the use of additional predictors is of value from a utility viewpoint only when they add additional variance explained in the criterion, beyond that which is accounted for by other, less expensive predictors (Schmidt and Hunter, 1998).

Various primary studies have examined whether SJTs significantly add to the prediction of job performance over cognitive ability, job knowledge, job experience, and personality (Chan and Schmitt, 2002; Clevenger *et al.*, 2001; Lievens *et al.*, 2005a; McDaniel *et al.*, 2001; Oswald *et al.*, 2004; Weekley and Jones, 1997, 1999). Recently, McDaniel *et al.* (2007) conducted a meta-analysis on the incremental validity of SJTs. They concluded that SJTs provide incremental validity over cognitive ability, varying from 3 percent to 5 percent. Furthermore, the incremental validity of SJTs over personality was estimated between 6 percent and 7 percent. Finally, the incremental validity of SJTs over both cognitive ability and personality ranged from 1 percent to 2 percent. Another recent study found that the incremental validity of SJTs over common predictors differed with respect to the criterion. For instance, SJTs explained incremental variance above cognitive ability (but not above personality) for predicting contextual performance (O'Connell *et al.*, 2007).

Although this research base attests to the incremental validity of SJTs over other predictors, a caveat is in order. McDaniel *et al.* (2007) warned that:

SJT correlations with job performance, cognitive ability, and the Big Five vary widely. One could clearly construct scenarios where SJTs could contribute substantially to a predictor composite or offer near zero incremental validity (McDaniel *et al.*, 2007, p. 83).

Construct-related validity

Over the years, various constructs have been linked to SJTs. According to Wagner and Sternberg (1985), the purpose of an SJT is to measure something other than academic intelligence (cognitive ability). They proposed that SJTs measure "tacit knowledge" or "practical intelligence" (i.e. practical know-how that is usually not openly expressed or stated and which must be acquired in the absence of direct instruction). Other research does not support this position and reveals that SJTs are related to cognitive ability (see also McDaniel and Whetzel, 2005). In the meta-analysis of McDaniel *et al.* (2001), it was found that SJTs show a correlation of 0.46 with cognitive ability, even though there was substantial variability around this estimate. For instance, video-based SJTs had lower correlations with cognitive ability than written SJTs (Weekley and Jones, 1997). Another example is that SJTs based on a job analysis were usually more highly related to cognitive ability than those not based on a job analysis (0.50 versus 0.38). Still other researchers posit that SJTs are alternative measures of job knowledge, job experience or interpersonal variables (McDaniel and Nguyen, 2001; Weekley and Jones, 1999).

Taken together, the extent to which SJTs tap different constructs seems to vary greatly. This is no surprise as SJT items may refer to a wide range of situations and include different types of content to which applicants must attend when making a decision. In addition, responses to SJT items with multiple options are the result of a combination of ability, experience, and personality. Recently, some efforts have been

undertaken to open the “black box” of what SJTs measure. Again, the type of response instructions mattered. Specifically, the meta-analysis of McDaniel *et al.* (2007) reported that SJTs with knowledge instructions correlated more highly with cognitive ability tests (0.35) than SJTs with behavioural tendency instructions (0.19). Conversely, SJTs with behavioural tendency instructions correlated more highly with Agreeableness (0.37), Conscientiousness (0.34), and Emotional Stability (0.35) than SJTs with knowledge instructions (0.19, 0.24, and 0.12, respectively). These results confirm that SJTs with knowledge instructions should be considered maximal performance measures, whereas SJTs with behavioural tendency instructions are typical performance measures.

Utility

As noted above, an important advantage of SJTs over more costly alternatives such as AC exercises is that SJTs can be used to test large groups of applicants at once (over the internet). Along these lines, recent research confirms the equivalence between written and internet-delivered SJTs (Ployhart *et al.*, 2003; Potosky and Bobko, 2004). Although this internet-based format is an important advantage, this does not address the economic utility of SJTs (in terms of monetary value) as compared to other selection procedures.

The economic utility of any selection procedure is among others determined by the criterion-related validity of the selection procedure and by the costs involved (Cronbach and Gleser, 1965). Unfortunately, no research has tested the economic utility of using SJTs. Therefore, we discuss the two main aspects of utility separately. With regard to validity, there is meta-analytic evidence that supports the criterion-related validity of SJTs (see above). In addition, the meta-analytic evidence about the incremental validity of SJTs over cognitive ability and personality might serve as another argument for the utility of SJTs in a selection battery.

With regard to the developmental costs of SJTs we contacted two SJT vendors and developers (e-mail communications, Joshua Sacco and Michael McDaniel, May 11, 2007). On average, the cost of developing a written SJT for a specific job for a specific organization ranged between \$60,00.00 and \$120,00.00. These broad cost estimates should be interpreted with caution. First, these developmental costs refer to job-specific SJTs. When an SJT is constructed for a specific job and for a specific organization, this tailored process will be more expensive than the one for generic SJTs. Second, the above estimates refer to SJTs developed from scratch. This means that all of the developmental stages that we described above are billed. Sometimes, some of the information needed (e.g. critical incidents) might already be available. Third, these cost estimates do not include ancillary studies (e.g. a study on the criterion-related validity of the SJT, norms based on applicant samples). Fourth, these developmental costs are for written SJTs. The costs are higher for video-based (multimedia) SJTs as these formats involve developing scripts, hiring actors, filming the performances, and editing the videos. In addition, the administration costs of video-based SJTs are higher as technological investments (VCR, PCs) have to be made for administering video-based and/or multimedia SJTs. Along these lines, McHenry and Schmitt (1994) warned that it is often necessary to double the original forecasted cost estimates associated with video-based (multimedia) tests.

Adverse impact

This question deals with the issue as to whether particular groups (e.g. White, male, and young applicants) systematically receive higher scores in SJTs? So far, only the effects of race and gender on SJT performance have been examined.

With respect to race, differences in mean SJT scores between racial subgroups are typically smaller than those reported for cognitive ability tests (Jensen, 1998). The meta-analysis of Nguyen *et al.* (2005a) found a difference in mean SJT scores between Whites and Blacks of 0.38 standard deviations in favour of White candidates. A key determinant of whether SJTs show adverse impact is the correlation of SJTs with cognitive ability. This correlation explained almost all of the variance in mean racial differences across studies. Thus, subgroup differences between Blacks and Whites are considerably reduced if SJTs measure primarily non-cognitive aspects of job performance. Additionally, video-based SJTs seem to result in less adverse impact than written SJTs because video-based SJTs are less cognitively loaded (Chan and Schmitt, 1997). Finally, SJTs with behavioural tendency instructions (measures of typical performance) showed lower adverse impact than SJTs with knowledge instructions (Nguyen *et al.*, 2005a). Given their lower adverse impact, SJTs are often used in a battery to find a trade-off between maximizing validity and reducing adverse impact (Pulakos and Schmitt, 1996). Yet, it should be noted that the lower reliability of SJTs might also partially explain the lower subgroup differences found.

With respect to gender, females seem to score slightly better than males on SJTs. In their meta-analysis, Nguyen *et al.* (2005a) found a difference in mean scores between females and males of 0.10 standard deviations in favour of females. This gender bias might be due to gender differences in terms of the personality traits triggered by the SJT situations. These scenarios are often interpersonal in nature. In general, females tend to score higher on traits such as Agreeableness or Sociability (Costa *et al.*, 2001).

Applicant perceptions

Generally, applicants prefer selection tools, which they perceive as job-related. That is one of the reasons why work samples and AC exercises typically receive favourable ratings (Hausknecht *et al.*, 2004). In this regard, it is not surprising that research on applicant reactions to SJTs showed that SJTs were perceived as favourable and that video-based formats even resulted in more positive perceptions than written formats (e.g. Chan and Schmitt, 1997). In addition, Richman-Hirsch *et al.* (2000) demonstrated that a multimedia SJT was seen as significantly more face valid, more enjoyable, and more modern than the computerized and written forms of the same SJT. Recently, Kanning *et al.* (2006) scrutinized applicant perceptions of SJT items that varied along interactivity, stimulus fidelity, and response fidelity. Interactive SJT items using videos in the stimulus and response component received the highest ratings.

Fakability

Given that SJTs are low-fidelity simulations and use a self-report format, it is relevant to examine the extent to which they are prone to deliberate response distortion (i.e. “faking good”). Hooper *et al.* (2006) summarized the available research evidence and discovered that differences in mean scores between respondents who were asked to respond as honestly as possible and respondents who were asked to “fake” varied

between 0.08 and 0.89 *SD*. They also concluded that the SJT faking effects are considerably smaller than in the case of personality measures.

Interestingly, Hooper *et al.* (2006) also identified several moderators that might make an SJT more fakable and that might explain the large differences across faking studies. First, when SJT items had a stronger cognitive loading, they were less fakable (see also Peeters and Lievens, 2005). Second, more transparent items were more fakable. Third, the type of response instructions was a key factor as it affected the cognitive loading and amount of response distortion in SJTs (Nguyen *et al.*, 2005b; Ployhart and Ehrhart, 2003). Behavioural tendency instructions were more fakable than knowledge-based instructions. Finally, the type of study design played a role. Laboratory findings were a worst-case scenario in comparison to real-life selection. Such experimental laboratory designs manipulate faking and investigate whether applicants can fake a test (i.e. ability to fake). This is not the same issue as whether applicants do fake a test in actual selection (i.e. motivation to fake).

Susceptibility to practice effects and coaching

When a selection procedure becomes popular, it can be assumed that candidates will attend commercial test coaching programs and adopt strategies to improve their test scores, thereby increasing their chances of being selected. This latter issue raises the question: Can SJT performance be enhanced through coaching?

Only one study has tackled this issue so far. Cullen *et al.* (2006) examined the coachability of SJTs developed for consideration as selection instruments in high-stakes testing (college admission process). Results indicated that some SJTs were susceptible to coaching. These results show that caution should be exerted with respect to the use of SJTs in high-stakes testing (e.g. admission, licensure, and accreditation exams).

A similar issue is whether candidates can improve their scores when they retake SJTs. Again, research is scarce. Lievens *et al.* (2005b) demonstrated that retest effects of SJTs were not larger than those of more traditional tests (cognitive ability). An important moderator seems to be the approach of constructing alternate SJT forms (Clause *et al.*, 1998; Oswald *et al.*, 2005). Specifically, Lievens and Sackett (2006b) compared various alternate form development approaches that differed in terms of the similarity of the items included in the alternate SJT forms. The approach that built in the least similarity among alternate SJT forms (i.e. random assignment of SJT items across forms) resulted in the smallest retest effects. However, this approach also produced a low correlation among the “alternate” forms (in the 0.30s).

Directions for future research

Although SJTs have established themselves as valid predictors in the employment and educational domain, we are only just starting to better understand them. As a first key avenue for future research, we need to enhance our understanding of why SJTs predict work behaviour. Prior research examined cognitive ability, experience, and personality as antecedents of SJT performance. This is only a start. Recently, procedural knowledge and implicit trait policies have been advocated as two plausible alternative explanations (Motowidlo *et al.*, 2006) for why SJTs are predictive of work behaviour. These might open a window of possibilities for more theory-based research on SJTs.

As a second key avenue for future research, we should examine the validity and utility of SJTs *vis-à-vis* other selection procedures. We need studies that investigate the incremental validity of SJTs over other low-fidelity simulations (situational interviews and behaviour description interviews). Similarly, the utility of SJTs over high-fidelity simulations (AC exercises) should be determined. In this comparative research it is important to keep the selection stage and the constructs measured constant. Only in that case, one might determine whether SJTs as measurement methods have added value over these other selection procedures.

A third critical gap in the extant research base is the need to understand how different SJT features impact on their effectiveness. Some initial steps have been undertaken on this route. As noted above, prior research (Chan and Schmitt, 1997; Lievens and Sackett, 2006; McDaniel *et al.*, 2007) has already identified the degree of stimulus fidelity (written vs video-based) and the type of response instructions (knowledge-based vs behavioural tendency) as key factors in determining the cognitive loading of SJTs. We need more studies that investigate the influence of other SJT features on adverse impact and validity. Some examples are the use of branched items, the type of subject matter experts, the level of item specificity, or the length of items. Generally, it would be helpful to construct a taxonomy of content domains of SJTs and examine how different content being captured by SJTs affects the relationship between SJT scores and external correlates (personality and cognitive ability).

Fourth, we welcome research that experiments with new stimulus and response formats for SJTs. A cartoon-based SJT is such an example of a new stimulus format. Although a cartoon-based format does not capture the wealth of information of video, it is much easier to administer through the Internet. An example of a new response format might consist of showing candidates a video-based situation and asking them to act out their response, while being videotaped by a camera or webcam. Future research should compare the effectiveness of these innovative formats to more traditional formats.

If SJTs really want to make inroads in international selection practice, a fifth critical research area is the cross-cultural transportability of SJTs (Lievens, 2006). That is, can SJTs developed in one culture be transported to and used as a valid predictor in another culture? In one of the sole studies on this topic, Such and Schmidt (2004) examined the validity of the same SJT in various countries. The SJT was valid in half of the countries, namely the UK and Australia. Conversely, it was not predictive in Mexico. The generalizability of SJTs to other contexts might be jeopardized if SJTs were used in a different context (e.g. job, organization, culture) and for a different criterion than intended. In cross-cultural applications of SJTs, tailoring the scoring key to the host culture might be a way of matching predictors and criteria. Research is needed to test this logic. So far, no studies have explored cultural differences in terms of the item stems, response options, or response option-construct linkages of SJTs.

Conclusion

This paper presented SJTs, including their characteristics, strengths, and weaknesses. Important strengths of SJTs are that they show criterion-related validity and incremental validity over cognitive ability and personality. Furthermore, applicant reactions are positive due to the job-relatedness of SJTs and SJTs have less adverse impact towards minorities than cognitive ability tests (if the cognitive loading of the

SJT is low). Finally, SJTs can be used to test large groups of applicants at once (over the Internet). In terms of weaknesses, SJTs might be prone to faking, practice, and coaching (although to a lesser extent than personality inventories). In addition, most SJTs are context-specific instruments, making it necessary to develop SJTs for specific jobs (job families) and cultures.

References

- Becker, T.E. (2005), "Development and validation of a situational judgment test of employee integrity", *International Journal of Selection and Assessment*, Vol. 13 No. 3, pp. 225-32.
- Bergman, M.E., Drasgow, F., Donovan, M.A. and Henning, J.B. (2006), "Scoring situational judgment tests: once you get the data, your troubles begin", *International Journal of Selection and Assessment*, Vol. 14 No. 3, pp. 223-35.
- Bruce, M.M. and Learner, D.B. (1958), "A supervisory practices test", *Personnel Psychology*, Vol. 11, pp. 207-16.
- Chan, D. and 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*, Vol. 82 No. 1, pp. 143-59.
- Chan, D. and Schmitt, N. (2002), "Situational judgment and job performance", *Human Performance*, Vol. 15 No. 3, pp. 233-54.
- Clause, C.C., Mullins, M.E., Nee, M.T., Pulakos, E.D. and Schmitt, N. (1998), "Parallel test form development: a procedure for alternative predictors and an example", *Personnel Psychology*, Vol. 51 No. 1, pp. 193-208.
- Clevenger, J., Pereira, G.M., Wiechmann, D., Schmitt, N. and Schmidt-Harvey, V.S. (2001), "Incremental validity of situational judgment tests", *Journal of Applied Psychology*, Vol. 86 No. 3, pp. 410-7.
- Cook, M. (2003), *Personnel Selection: Adding Value through People*, Wiley, Chichester.
- Costa, P.T., Terracciano, A. and McCrae, R.R. (2001), "Gender differences in personality traits across cultures: robust and surprising findings", *Journal of Personality and Social Psychology*, Vol. 81 No. 2, pp. 322-31.
- Cronbach, L.J. and Gleser, G.C. (1965), *Psychological Tests and Personnel Decisions*, University of Illinois Press, Urbana, IL.
- Cullen, M.J., Sackett, P.R. and Lievens, F. (2006), "Threats to the operational use of situational judgment tests in the college admission process", *International Journal of Selection and Assessment*, Vol. 14 No. 2, pp. 142-55.
- Dalessio, A.T. (1994), "Predicting insurance agent turnover using a video-based situational judgment test", *Journal of Business and Psychology*, Vol. 9 No. 1, pp. 23-32.
- Funke, U. and Schuler, H. (1998), "Validity of stimulus and response components in a video test of social competence", *International Journal of Selection and Assessment*, Vol. 6, pp. 115-23.
- Hausknecht, J.P., Day, D.V. and Thomas, S.C. (2004), "Applicant reactions to selection procedures: an updated model and meta-analysis", *Personnel Psychology*, Vol. 57 No. 3, pp. 639-83.
- Hooper, A.C., Cullen, M.J. and Sackett, P.R. (2006), "Operational threats to the use of SJTs: faking, coaching, and retesting issues", in Weekley, J.A. and Ployhart, R.E. (Eds), *Situational Judgment Tests: Theory, Measurement and Application*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 205-32.

-
- Hunter, D.R. (2003), "Measuring general aviation pilot judgment using a situational judgment technique", *International Journal of Aviation Psychology*, Vol. 13 No. 4, pp. 373-86.
- Jensen, A.R. (1998), *The G Factor: The Science of Mental Ability*, Praeger, Westport, CT.
- Kanning, U.P., Grewe, K., Hollenberg, S. and Hadouch, M. (2006), "From the subjects' point of view – reactions to different types of situational judgment items", *European Journal of Psychological Assessment*, Vol. 22 No. 3, pp. 168-76.
- Konradt, U., Hertel, G. and Joder, K. (2003), "Web-based assessment of call center agents: development and validation of a computerized instrument", *International Journal of Selection and Assessment*, Vol. 11 Nos 2-3, pp. 184-93.
- Lievens, F. (2006), "International situational judgment tests", in Weekley, J.A. and Ployhart, R.E. (Eds), *Situational Judgment Tests: Theory, Measurement and Application*, Laurence Erlbaum Associates, Mahwah, NJ, pp. 279-300.
- Lievens, F., Buyse, T. and Sackett, P.R. (2005a), "The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains", *Journal of Applied Psychology*, Vol. 90 No. 3, pp. 442-52.
- Lievens, F., Buyse, T. and Sackett, P.R. (2005b), "Retest effects in operational selection settings: development and test of a framework", *Personnel Psychology*, Vol. 58 No. 4, pp. 981-1007.
- Lievens, F. and Sackett, P.R. (2006a), "Video-based versus written situational judgment tests: a comparison in terms of predictive validity", *Journal of Applied Psychology*, Vol. 91 No. 5, pp. 1181-8.
- Lievens, F. and Sackett, P.R. (2006b), "Situational judgment tests in high stakes settings: issues and strategies with generating alternate forms", *Journal of Applied Psychology*, Vol. 92 No. 4, pp. 1043-55.
- McClough, A.C. and Rogelberg, S.G. (2003), "Selection in teams: an exploration of the teamwork knowledge, skills, and ability test", *International Journal of Selection and Assessment*, Vol. 11 No. 1, pp. 56-66.
- McDaniel, M.A. and Nguyen, N.T. (2001), "Situational judgment tests: a review of practice and constructs assessed", *International Journal of Selection and Assessment*, Vol. 9 Nos 1-2, pp. 103-13.
- McDaniel, M.A. and Whetzel, D.L. (2005), "Situational judgment test research: informing the debate on practical intelligence theory", *Intelligence*, Vol. 33 No. 5, pp. 515-25.
- McDaniel, M.A., Hartman, N.S., Whetzel, D.L. and Grubb, W.L. (2007), "Situational judgment tests, response instructions, and validity: a meta-analysis", *Personnel Psychology*, Vol. 60 No. 1, pp. 63-91.
- McDaniel, M.A., Morgeson, F.P., Finnegan, E.B., Campion, M.A. and Braverman, E.P. (2001), "Predicting job performance using situational judgment tests: a clarification of the literature", *Journal of Applied Psychology*, Vol. 86 No. 4, pp. 730-40.
- McHenry, J.J. and Schmitt, N. (1994), "Multimedia testing", in Rumsey, M.G. and Walker, C.B. (Eds), *Personnel Selection and Classification*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 193-232.
- Morgeson, F.P., Reider, M.H. and Campion, M.A. (2005), "Selecting individuals in team settings: the importance of social skills, personality characteristics, and teamwork knowledge", *Personnel Psychology*, Vol. 58 No. 3, pp. 583-611.
- Motowidlo, S., Dunnette, M.D. and Carter, G.W. (1990), "An alternative selection procedure: the low-fidelity simulation", *Journal of Applied Psychology*, Vol. 75 No. 6, pp. 640-7.

- Motowidlo, S.J., Hooper, A.C. and Jackson, H.L. (2006), "Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items", *Journal of Applied Psychology*, Vol. 91 No. 4, pp. 749-61.
- Nguyen, N.T., Biderman, M.D. and McDaniel, M.A. (2005b), "Effects of response instructions on faking a situational judgment test", *International Journal of Selection and Assessment*, Vol. 13 No. 4, pp. 250-60.
- Nguyen, N.T., McDaniel, M.A. and Whetzel, D.L. (2005a), "Subgroup differences in situational judgment test performance: a meta-analysis", paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA, April.
- O'Connell, M.S., Hartman, N.S., McDaniel, M.A., Grubb, W.L. and Lawrence, A. (2007), "Incremental validity of situational judgment tests for task and contextual job performance", *International Journal of Selection and Assessment*, Vol. 15 No. 1, pp. 19-29.
- Olson-Buchanan, J.B., Drasgow, F., Moberg, P.J., Mead, A.D., Keenan, P.A. and Donovan, M.A. (1998), "Interactive video assessment of conflict resolution skills", *Personnel Psychology*, Vol. 51 No. 1, pp. 1-24.
- Oswald, F.L., Friede, A.J., Schmitt, N., Kim, B.K. and Ramsay, L.J. (2005), "Extending a practical method for developing alternate test forms using independent sets of items", *Organizational Research methods*, Vol. 8 No. 2, pp. 149-64.
- Oswald, F.L., Schmitt, N., Kim, B.H., Ramsay, L.J. and Gillespie, M.A. (2004), "Developing a biodata measure and situational judgment inventory as predictors of college student performance", *Journal of Applied Psychology*, Vol. 89 No. 2, pp. 187-208.
- Peeters, H. and Lievens, F. (2005), "Situational judgment tests and their predictiveness of college students' success: the influence of faking", *Educational and Psychological Measurement*, Vol. 65 No. 1, pp. 70-89.
- Ployhart, R.E. (2006), "Staffing in the 21st century: new challenges and strategic opportunities", *Journal of Management*, Vol. 32 No. 6, pp. 868-97.
- Ployhart, R.E. and 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*, Vol. 11 No. 1, pp. 1-16.
- Ployhart, R.E., Porr, W. and Ryan, A.M. (2004), "New development in SJTs: scoring, coaching and incremental validity", paper presented at the 19th annual convention of the Society for Industrial and Organizational Psychology, Chicago, IL, April.
- Ployhart, R.E., Weekley, J.A., Holtz, B.C. and Kemp, C.F. (2003), "Web-based and paper-and-pencil testing of applicants in a proctored setting: are personality, biodata, and situational judgment tests comparable?", *Personnel Psychology*, Vol. 56 No. 3, pp. 733-52.
- Potosky, D. and Bobko, P. (2004), "Selection testing via the internet: practical considerations and exploratory empirical findings", *Personnel Psychology*, Vol. 57 No. 4, pp. 1003-34.
- Pulakos, E.D. and Schmitt, N. (1996), "An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity", *Human Performance*, Vol. 9 No. 3, pp. 241-58.
- Richardson, Bellows, Henry & Co. (1981), *Technical Reports Supervisory Profile Record*, Richardson, Bellows, Henry & Co., Washington, DC.
- Richman-Hirsch, W.L., Olson-Buchanan, J.B. and Drasgow, F. (2000), "Examining the impact of administration medium on examinee perceptions and attitudes", *Journal of Applied Psychology*, Vol. 85 No. 6, pp. 880-7.

-
- Ryan, A.M., McFarland, L., Baron, H. and Page, R. (1999), "An international look at selection practices: nation and culture as explanations for variability in practice", *Personnel Psychology*, Vol. 52 No. 2, pp. 359-91.
- Smiderle, D., Perry, B.A. and Cronshaw, S.F. (1994), "Evaluation of video-based assessment in transit operator selection", *Journal of Business and Psychology*, Vol. 9 No. 1, pp. 3-22.
- Schmidt, F.L. and 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*, Vol. 124 No. 2, pp. 262-74.
- Schmitt, N. and Chan, D. (2006), "Situational judgment tests: method or construct?", in Weekley, J. and Ployhart, R.E. (Eds), *Situational Judgment Tests*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 135-56.
- Stevens, M.J. and Campion, M.A. (1999), "Staffing work teams: development and validation of a selection test for teamwork", *Journal of Management*, Vol. 25 No. 2, pp. 207-28.
- Such, M.J. and Schmidt, D.B. (2004), "Examining the effectiveness of empirical keying: a cross-cultural perspective", paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL, April.
- Wagner, R.K. and Sternberg, R.J. (1985), "Practical intelligence in real world pursuits: the role of tacit knowledge", *Journal of Personality and Social Psychology*, Vol. 49 No. 2, pp. 436-58.
- Weekley, J.A. and Jones, C. (1997), "Video-based situational testing", *Personnel Psychology*, Vol. 50 No. 1, pp. 25-49.
- Weekley, J.A. and Jones, C. (1999), "Further studies of situational tests", *Personnel Psychology*, Vol. 52 No. 3, pp. 679-700.
- Weekley, J.A. and Ployhart, R.E. (2006), *Situational Judgment Tests: Theory, Measurement and Application*, Jossey-Bass, San Francisco, CA.

About the authors

Filip Lievens received his PhD from Ghent University, Belgium and is currently Professor at the Department of Personnel Management and Work and Organizational Psychology at Ghent University. His research interests include organizational attractiveness, high-stakes testing, and alternative selection procedures (assessment centres and situational judgment tests). Filip Lievens is the corresponding author and can be contacted at: filip.lievens@ugent.be

Helga Peeters received her PhD from Ghent University, Belgium. She currently works as HR Research Expert at Securex Research Center, Belgium.

Eveline Schollaert received her Master's degree in 2006 from Ghent University, Belgium. She is currently a teaching assistant at the Department of Personnel Management and Work and Organizational Psychology at Ghent University, Belgium. Her research interests include personnel selection and competency modeling.