Situational Judgment Tests

Filip Lievens

Department of Personnel Management, Work and Organizational Psychology,

Ghent University, Belgium

Address: Henri Dunantlaan 2, 9000 Ghent, Belgium.

Phone: +32(0)92646453

Fax: +32(0)92646454

E-mail: filip.lievens@ugent.be

Britt De Soete

Department of Personnel Management, Work and Organizational Psychology,

Ghent University, Belgium

Address: Henri Dunantlaan 2, 9000 Ghent, Belgium.

Phone: +32(0)92646459

Fax: +32(0)92646454

E-mail: Britt.DeSoete@UGent.be

Lievens, F., De Soete, B. (2015). Situational Judgment Test. In: James D. Wright (editor-in-chief), International Encyclopedia of the Social & Behavioral Sciences (pp. 13–19), 2nd edition, Vol 22. Oxford: Elsevier.

Keywords

Situational Judgment Tests, personnel selection, human resource management, simulations, reliability, construct-related validity, criterion-related validity, incremental validity, subgroup differences, test-taker perceptions, stimulus format, response format, response instruction, prediction, diversity, procedural knowledge

Abstract

In this chapter, we give an overview of situational judgment tests (SJTs) as selection instruments. Their history, basic characteristics, and development are presented. The available research evidence regarding their reliability, construct-related validity, criterion-related validity, incremental validity, subgroup differences, and test-taker perceptions is also reviewed. As a general conclusion, the increasing popularity of SJTs in personnel selection seems to be accredited to their potential to capture a variety of constructs and for different purposes. Additionally, SJTs are able to predict several job-related and/or academic criteria while at the same time offering prospects permitting to select for diversity.

Reference list (for further reading)

Arthur, W., & 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.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143-159.

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.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410-417.

Sackett, P.R., & Lievens, F. (2008). Personnel selection. Annual Review of Psychology, 59, 419-45.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). 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, 90*, 442-452.

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

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: a review of recent research. *Personnel Review, 37*, 426-441.

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97*, 460-468.

Lievens, F., De Corte, W., & Westerveld, L. (in press). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management*.

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

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

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowlegde from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95*, 321-333.

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.

Oswald, F. L., Schmit, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*, 187-207.

Weekley, J, & Ployhart, R.E. (2006). *Situational judgment tests*. San Francisco: Jossey-Bass.

Whetzel, D.L. & McDaniel, M.A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*, 188-202.

Situational Judgment Tests

History and Definition

During World War II (WWII), military selection psychologists were in need of a tool to select competent soldiers to join the armed forces. They developed a job test that consisted of detailed and realistic descriptions of challenging military situations. All descriptions were situations that armed forces were likely to encounter while on the job. After reading each situation, recruits were presented with several potential reactions to the given threat or challenge and they were asked which reaction they considered the most effective response (Northrop, 1989). The instrument turned out to be a success. On the one hand, it gave a realistic job preview of what was to come, thereby discouraging recruits with an unfavorable person-organization fit and lowering attrition rates for the army. On the other hand, the tool enabled to measure recruits' judgment skills in job-related settings, thereby significantly facilitating competent new soldier selection. The aforementioned instrument can be considered one of the first situational judgment tests (SJTs). After WWII, several similar tests were designed to capture supervisory potential (e.g., Bruce, 1974; Cardall, 1942; File, 1945; Greenberg, 1963). In 1990, Motowidlo and colleagues framed the SJT as a new alternative measurement procedure for personnel selection (Motowidlo, Dunnette, & Carter, 1990) and thereby reinvigorated interest in SJTs among scientists and practitioners. SJTs present test-takers with realistic job situations, followed by potential response options out of which candidates have to select the most appropriate response (Motowidlo et al., 1990). SJTs are considered measurement tools that aim to capture job-related competencies and skills (Lievens, Peeters, & Schollaert, 2007). Figure 1 shows an example SJT item.

< Insert Figure 1 about here >

Similar to assessment centers and work samples, SJTs are simulation-based instruments. Simulations are based on the behavioral consistency logic (Lievens & De Soete,

2012). That is, the assumption that candidates' performance during the selection procedure will be consistent with their future performance on the job. The difference between SJTs and other simulation-based measurement instruments lies in their level of fidelity. Fidelity can be defined as the extent to which the selection procedure mirrors the actual job situation (Callinan & Robertson, 2000). As assessment centers and work samples require actual behavior during the selection phase, such instruments can be considered high-fidelity simulations. High-fidelity simulations are more expensive to administer and are therefore generally used in small samples, during later selection stages. In contrast, SJTs are described as low-fidelity simulations (Motowidlo et al., 1990) as they do not require test-takers to display actual behavior but instead confront them with written descriptions of realistic job situations. As a result, SJTs can be administered to large applicant groups in preliminary selection stages. Since their reintroduction by Motowidlo and colleagues in 1990, SJTs have become attractive selection instruments for practitioners who are looking for cost-effective instruments for measuring a wide variety of predominantly interpersonally-oriented constructs.

SJT Development

The development of SJTs typically consists of three stages (Motowidlo et al., 1990). In the first stage, the stimulus material is developed. It is advisable to start this phase with a job analysis, so that the knowledge, skills, abilities and other characteristics (KSAOs) are identified which are considered to be crucial for job performance. For each of these competencies, critical incidents of work situations are gathered from subject matter experts (i.e., incumbents, their supervisors, clients) or from archival sources. As a next step, test developers select the best non-redundant critical incidents from the total pool and rewrite them into test items of similar length and format.

Second, the response options are developed. To this end, all items are presented to a different group of subject matter experts or to inexperienced employees, which are asked to formulate possible responses to the given job situations. In this stage, test developers aim to

collect a satisfactory amount of responses for every single item, and these responses should capture a wide range of effectiveness.

As a third and last step, the SJT's scoring key is developed. Test developers mostly opt for a rational or an empirical scoring key. When developing a rational scoring key, subject matter experts are asked to either identify the best and the worst response options or to rate all responses per item on their effectiveness. Unlike the rational scoring key, an empirical scoring key does not use expert judgments. Instead, the SJT is administered among a large pool of incumbents, whose responses are then linked to a criterion (e.g., job performance). SJT responses that are mostly chosen by high performing employees are labeled as 'correct' or 'highly effective', whereas the opposite takes place for SJT responses that are not endorsed by high performing employees or selected by low performing individuals.

SJT Design Considerations

Although most SJTs are designed according to the predefined steps described above, test developers and their clients also face various design decisions, which may impact on the specific outlook of the SJT developed. These decisions are typically driven by the available resources and the objectives (hiring, promotion, training, recruitment, etc.) one wants to accomplish with the SJT.

A first set of decisions concerns the characteristics of the item. A first aspect of the item stem that varies between SJTs is the *item length*. Some SJTs provide test-takers with a simple and short situational description (e.g., "A colleague in your team dodges her duties and you have to take over a lot of her tasks."), whereas others present the participant with detailed descriptions of the job situation and its background. Similarly, items may vary in their level of *contextualization*. Some SJTs are built to measure KSAOs for specific jobs, whereas others are meant to measure generic competencies or skills across several jobs on the same level. As a result, SJTs in the former category will often be characterized by a higher degree of contextualization than the latter types. Highly contextualized SJT items provide specified information about the organizational setting and job context, which is illustrated by mentioning job-specific equipment or by using job terminology. A final item-related difference

between SJTs is their level of *interactivity*. Whereas traditional SJTs present each test-taker with the same (sequence of) items, interactive or so-called 'branched' or 'nested' SJTs take into account the test-taker's response to former items to decide which items will be presented subsequently (Olson-Buchanan et al., 1998). Accordingly, test-takers are confronted with the consequences of their response actions. For example, if a participant decides to fire an unproductive employee in one item, the next item may deal with an uproar in the team as a consequence of the dismissal. The main advantage of these branched SJTs, is that they permit to mirror the dynamics of an actual interaction, while maintaining a certain degree of standardization (Lievens et al., 2007).

Design decisions should also be made regarding the SJT's *stimulus and response modality*. Modality refers to the way the information is presented to the candidate (for stimuli) or the manner in which the candidates are required to answer (for responses). Regarding stimulus modality, SJTs traditionally have a text-based format that presents test-takers with written descriptions of job situations. As calls have been made to develop more realistic measurement instruments (e.g., Lane & Stone, 2006), i.e., instruments that show a greater resemblance with the actual job for which they are selecting applicants, written SJT stimuli can be replaced by video-based or multimedia stimuli (Chan & Schmitt, 1997; Weekley & Jones, 1997). In such SJTs, test-takers are presented with short videos of job-related situations. At a critical point, the video freezes and the test-taker is subsequently required to select the most appropriate response to the presented situation.

Regarding response modality, SJTs traditionally have a text-based multiple choice format, which present test-takers with a predetermined set of written responses. However, recently also alternative response modalities have made their entrance in the SJT domain. For instance, Kanning, Grewe, Hollenberg, and Hadouch (2006) experimented with video-based instead of written response options. More recently, Crook and colleagues introduced the single-response SJT (Crook at al., 2011; Motowidlo, Crook, Kell, & Naemi, 2007). Single-response SJTs confront test-takers with only one response option, which they have to rate on its effectiveness by means of a Likert scale. The main advantage of single-response SJTs

is their less labor-intensive test design. In contrast to traditional multiple choice SJTs, single-response SJTs require only one group of subject matter experts for test design, which are asked to come up with critical incidents that struck them as extremely effective or ineffective. Based on this information, both item stems and responses can be derived, and the second development stage of traditional SJTs can be skipped.

A third decision concerns the SJT's *response instructions*. Both knowledge-based and behavioral-based response instructions are frequently used (McDaniel, Hartman, Whetzel, & Grubb III, 2007). Knowledge-based response instructions ask the candidate to display their knowledge of the responses' effectiveness by selecting the best/worst response option ("What is the best/worst answer?") or by rating each response option on its perceived effectiveness. As a result, SJTs with knowledge-based instructions are measures of maximal performance. In contrast, SJTs with behavioral-based response instructions ask candidates to report how they would respond in the presented situation ("What are you most likely to do?") and are therefore considered as typical performance measures.

Finally, a fourth set of decisions deals with the SJT's *scoring key*. As has been mentioned earlier, empirical and rational scoring keys are the most common choices for scoring keys. Therefore, hybrid forms have also been developed. For instance, an extant empirical scoring key is given to experts to make sure that "it makes sense". Alternatively, one might start with an expert-scoring key and later on validate it empirically.

An Evidence-based Evaluation of SJTs

Since Motowidlo launched the SJT as an alternative measurement procedure for personnel selection in the early nineties (Motowidlo et al., 1990), research on SJTs has made great strides forward. Today, over 60 studies have been devoted to the development and psychometric properties of SJTs. On the basis of the available research, we review below SJTs in terms of their reliability, construct-related validity, criterion-related validity, incremental validity, subgroup differences, and test-taker perceptions.

Are SJT Scores Reliable?

To assess whether or not SJTs scores are reliable, most prior research has focused on internal consistency reliability. For example, a meta-analysis of McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001), based on 39 different SJTs and over 10,000 participants, found internal consistency reliability coefficients ranging from .43 to .94. A more recent study of Catano, Brochu, and Lamerson (2012) revealed a corrected weighted mean internal consistency of .46 based on 56 coefficients.

The variety in internal consistency reliability coefficients is in the first place a function of the instrument's length. SJT scores based on more items generally demonstrate higher internal consistency reliability. In addition, Ployhart and Ehrhart (2003) found that response instructions may serve as a second driver of internal consistency variability in SJTs. More specifically, SJT response instructions requiring the participant to rate each response on its effectiveness by means of a Likert-like scale resulted in the highest internal consistency reliability coefficients. Instructions that asked participants to choose two response options (e.g., "What is your most/least likely response?" or "What is the most/least effective response?") displayed somewhat lower internal consistency reliability coefficients, and instructions requiring participants to select one single response option (e.g., "What is your most likely response?" or "What is the most effective response?") resulted in the lowest internal consistency reliability coefficients.

Besides test length and response instructions, the often observed heterogeneous nature of many SJTs is probably one of the main reasons for their low internal consistency reliability (Lievens et al., 2007). As most SJTs are heterogeneous at the item level and aim to capture a plethora of constructs, internal consistency may therefore not be the most preferred means to evaluate SJT scores' reliability. Instead, researchers have proposed that alternate-form reliability or test-retest reliability may be more appropriate in the case of SJTs (Whetzel & McDaniel, 2009). Unfortunately, both of these reliabilities are rarely reported. One recent exception is the study of Catano and colleagues (2012) which reported test-retest reliabilities ranging from .66 to .82.

What Do SJTs Measure?

As SJTs have been used to capture a wide variety of constructs, answering this question is a challenge. Over the past years, SJTs have been developed to measure several competencies as diverse as entry-level managerial skills (Motowidlo et al., 1990), leadership skills (Oostrom, Born, Serlie, & Van der Molen, 2012), team work skills (Prewett, Brannick, & Peckler, 2013), emotional intelligence (Libbrecht & Lievens, 2012; Libbrecht, Lievens, Carette, & Côté, in press; Sharma, Gangopadhyay, Austin, & Mandal, 2013), interpersonal skills of medical students (Lievens & Sackett, 2012), aviation pilot judgment (Hunter, 2003), personal initiative (Bledow & Frese, 2008), and integrity (Becker, 2005; De Meijer, Born, Van Zielst, Van der Molen, 2010).

In an attempt to develop a content-based typology, Christian, Edwards, and Bradley (2010) identified and clustered all construct domains assessed by SJTs. They found that about 70% of the extant SJTs aim to capture either leadership or interpersonal skills. To assess which factors determine SJT performance, several meta-analyses have demonstrated a relatively high correlation between SJT performance on the one hand and cognitive ability measures or personality traits on the other hand. In terms of SJTs scores' relation with cognitive ability scores, meta-analyses reveal correlation coefficients ranging from r = .32 (McDaniel et al., 2001) to r = .46 (McDaniel et al., 2007). Four moderating factors for the relationship between SJT performance and cognitive ability can be derived from the literature. First, SJTs that are developed on the basis of a thorough job-analysis seem to display higher correlations with cognitive ability than SJTs that do not start from a job analysis (McDaniel et al., 2001). Second, the more detailed the SJT question, the lower the SJT's correlation with cognitive ability appears to be (McDaniel et al., 2001). Third, several studies have emphasized the importance of the stimulus format for the constructs assessed. For instance, SJTs with a video-based stimulus format demonstrate a lower correlation with performance on cognitive instruments than paper-and-pencil SJTs (Chan & Schmitt, 1997; Lievens & Sackett, 2006). Finally, in their meta-analysis in 2007, McDaniel et al. (2007) identified the response instructions as the fourth important influencer of cognitive

saturation. That is, SJTs with knowledge-based response instructions display higher correlations with cognitive ability than SJTs with behavioral tendency instructions.

SJT scores have also found to be correlated with personality. McDaniel et al. (2007) demonstrated that a SJT's relation with personality traits is a function of its response instructions. That is, SJTs with behavioral response instructions show significantly higher correlations with Agreeableness (r = .37), Conscientiousness (r = .34), and Emotional Stability (r = .35) as compared to SJTs with knowledge-based response instructions (r = .19, r = .24, and r = .12, respectively).

What Do SJTs Predict?

One of the most important evaluation criteria when deciding whether or not to use selection instruments refers to the extent to which they are related to job-related performance domains. Meta-analytic research has confirmed the expectation that SJTs are valuable predictors of job performance. A first meta-analysis on the criterion-related validity of SJT scores revealed a validity coefficient of r = .34 for predicting job performance (McDaniel et al., 2001). As one of the most important moderators of criterion-related validity, McDaniel and colleagues demonstrated the presence or absence of a job analysis to be an influential factor. SJTs that were based on a thorough job analysis displayed higher validity than SJTs that did not use a job analysis as their starting point. Several years later, a second meta-analysis was undertaken, which included more data than the former and which revealed an estimated population criterion-related validity coefficient of r = .26 (McDaniel et al., 2007). The most recent meta-analysis on criterion-related validity of SJTs thus far was published in 2010. On the basis of 84 studies, Christian and colleagues (2010) found validity coefficients of r = .38 for SJTs measuring teamwork, r = .28 for SJTs on leadership, and r = .25 for SJTs capturing interpersonal skills. Additionally, two extra moderators of SJT criterion-related validity were identified. First, the importance of careful predictor-criterion matching was emphasized as SJTs measuring specific competencies (e.g., interpersonal skills) demonstrated the highest validity for predicting matching criteria domains (e.g., interpersonal job performance, see also Lievens, Buyse, & Sackett, 2005). Second, it was found that a

SJT's stimulus modality may influence its criterion-related validity. More specifically, video-based SJTs displayed higher criterion-related validity coefficients than text-based SJTs for predicting interpersonal skills (see also Lievens & Sackett, 2006).

Apart from showing sufficient criterion-related validity in employment settings, SJT scores have also found to be useful for predicting academic success. For example, Lievens and colleagues repeatedly demonstrated that SJTs are good predictors of academic performance among student physicians (Lievens, 2013; Lievens et al., 2005; Lievens & Sackett, 2006, 2012). In a similar vein, Oswald, Schmitt, Kim, Ramsay, and Gillespie (2004) developed a SJT for predicting student performance. Their SJT aimed to capture 12 dimensions of college student performance and was successful in predicting academic success.

Finally, SJT scores have demonstrated consistent incremental validity over and above scores on more established selection instruments (Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt Harvey, 2001). This implies that SJTs succeed in predicting unique variance in criterion performance, which cannot be accounted for by other predictors. Various studies have demonstrated incremental validity of SJTs for predicting job or academic performance over cognitive ability, job knowledge, job experience, and personality (Chan & Schmitt, 2002; Clevenger et al., 2001; Lievens & Patterson, 2011; McDaniel et al., 2001; McDaniel et al., 2007; O'Connell, Hartman, McDaniel, Grub III, & Lawrence, 2007; Oswald et al., 2004; Weekley & Jones, 1997, 1999). In their meta-analysis, McDaniel et al. (2007) estimated the incremental validity of SJTs over cognitive ability between 3 and 5 percent, with somewhat higher incremental validity coefficients when behavioral instead of knowledge-based response instructions were used. Additionally, it was found that SJTs provide incremental validity over personality, varying from 6 to 7 percent, with the highest incremental validity for SJTs with knowledge-based response instructions. Over a combination of cognitive ability and personality, SJTs provided incremental validity of 1 to 2 percent. Besides response instructions, also the criterion type has been proven to be an influential factor of incremental validity variability (O'Connell et al., 2007). For instance,

O'Connell and colleagues found incremental validity for SJTs over cognitive ability but not over personality for predicting contextual performance, whereas this was not the case for predicting task performance. More research is needed to confirm this finding on a more general level, though.

Recently, conceptual progress has been made to explain why SJTs are valid predictors. According to Motowidlo and Beier (2010), SJTs measure general as well as specific knowledge of the costs and the consequences of job-related actions. General knowledge refers to implicit knowledge about the relationships between expressions of personality traits and effective job performance (i.e., implicit trait policies). That is, such general knowledge refers to people's judgments about the costs and benefits of engaging in courses of action as response to SJT situations. Specific knowledge is different from general knowledge because it is limited to specific job situations, which may include exceptions to successful trait policies (e.g., specific job situations where assertive reactions are more valued than agreeable reactions). The combination of both knowledge types captured by SJTs is assumed to be predictive of effective performance.

What About SJTs and Diversity?

An additional reason for the popularity of SJTs in the field of personnel selection is that they are introduced as "alternative" predictors, which might display smaller subgroup differences than cognitive ability tests. So far, the research evidence is a bit more mixed. Ethnic subgroup differences on SJTs have varied from approximately one standard deviation (Chan & Schmitt, 1997) to almost zero (Olson-Buchanan, et al., 1998), with Caucasians obtaining higher scores than Black (d = .38), Hispanic (d = .24), Asian (d = .29), and European minority (d = .38) test-takers (De Meijer 2008; Whetzel, McDaniel, & Nguyen, 2008).

Some factors may explain the great variety in subgroup differences assonated with SJT scores. First, meta-analytic research identified cognitive loading as one of the most important drivers of ethnic subgroup differences in SJT performance. Cognitive loading refers to the extent that SJT performance correlates with performance on a cognitive ability test.

Similar to assessment centers and work samples (Bobko, Roth, & Buster, 2005; Dean, Bobko, & Roth, 2008; Roth, Bobko, McFarland, & Buster, 2008), SJTs with a higher cognitive loading display substantially larger ethnic subgroup differences (Roth, Bobko, & Buster, 2013; Whetzel et al., 2008). Second, the personality loading, i.e., the correlation between the SJT and each of the Big Five personality factors, has been identified as a smaller driver of ethnic subgroup differences in SJTs, so that Black-White and Asian-White differences in SJT performance are smaller when the SJT displays a higher correlation with emotional stability and Hispanic-White differences are smaller when the SJT displays a higher correlation with conscientiousness and agreeableness (Whetzel et al., 2008). Third, SJT response instructions have been shown to influence ethnic performance differences. Knowledge-based response instructions ("What is the best response?") generally lead to larger ethnic subgroup differences than behavioral tendency response instructions ("How would you respond?"), which is most likely due to the greater cognitive loading of SJTs with knowledge-based response instructions (Whetzel et al., 2008). A fourth moderator concerns the SJT response process. SJTs that require candidates to rate each response option's effectiveness by means of a Likert scale, are more susceptible to ethnic subgroup differences because SJT performance is influenced by the White-Black difference in extreme scoring preferences. By controlling for elevation and scatter in responses, Black-White performance differences on this SJT type are substantially reduced. In addition, this correction is particularly promising as it simultaneously improves validity (McDaniel, Psotka, Legree, Powell Yost, & Weekley, 2011).

In terms of gender differences in SJT performance, a meta-analysis demonstrated that female test-takers on average perform slightly higher than male test-takers (d = -.11; Whetzel et al., 2008). Larger female advantages occur when the SJT is more correlated with agreeableness and conscientiousness (Weekley, Ployhart, & Harold, 2004; Whetzel, et al., 2008).

Do Candidates Like SJTs?

A last aspect for evaluating SJTs relates the test perceptions. In general, simulation instruments, such as assessment centers, work samples, and SJTs, are more favorably received by applicants than cognitive ability tests, personality inventories, biodata, and integrity tests (Hausknecht et al., 2004; Oostrom, Born, Serlie, & Van der Molen, 2010). Some researchers have identified specific SJT factors which may influence test perceptions or attitudes. One of the most important drivers of test perceptions in SJTs thus far, seems to be the stimulus modality. Richman-Hirsch, Olson-Buchanan, and Drasgow (2000) examined the effect of stimulus modality on test perceptions and attitudes for three content-wise identical conflict management SJTs. The first SJT had a paper-and-pencil format. The second SJT was identical to the first but used a computer screen and an automatic page turner to display the information. The third SJT presented test-takers with exactly the same scenarios but in video format. Applicants watched videos of job-related conflict situations and were asked to select the best response out of a set of four written options. Results demonstrated that test-takers reported significantly more favorable face validity perceptions and test attitudes for the video SJT as compared to the other two formats. Furthermore, there was no difference in reported perceptions and attitudes between the paper-and-pencil SJT and the computerized version, which suggests that computerizing test content is not sufficient to influence test perceptions. Along the same lines, Chan and Schmitt (1997) compared a written SJT with a content-wise identical video variant. They found significantly higher face validity perceptions for the latter SJT. Finally, Kanning et al. (2006) examined which SJT factors improve test perceptions. They discovered that changing the stimulus or response modality from written to video format alone was not enough to increase test perceptions, but that interactive SJTs with video stimuli as well as video response options received the most favorable test perceptions.

Epilogue

In the last two decades, SJTs have become a popular selection instrument that is widely used by practitioners and intensively studied by researchers. Their popularity in the field of personnel selection can largely be accredited to the potential of SJTs to capture a

variety of constructs, in diverse settings, and for different purposes. Additionally, research has repeatedly proven that SJTs are able to predict several job-related and/or academic criteria while at the same time offering prospects permitting to select for diversity.

Moreover, the future of SJTs for research and practice is looking bright. At the risk of being self-promoting, below we give a brief overview of some of our own research priorities in the next years. One avenue that we see as increasingly important is the use of SJTs in cross-cultural settings. To this end, research is needed to examine the intercultural transportability and robustness of SJTs (see Lievens, 2006, for an overview). Second, calls have been made to increase our conceptual understanding of the different components of SJTs and their effects on key selection outcomes. Here we advocate the use of a building block approach (e.g., Arthur & Villado, 2008; Lievens et al., in press). That is, instead of treating SJTs as holistic entities, the SJT method can be conceptualized as a combination of various predictor method factors (i.e., response instructions, stimulus format, response format, etc.). By keeping other factors constant, researchers are able to identify the impact of specific SJT factors on the criterion of interest (e.g., construct-related validity, criterion-related validity, subgroup differences, Chan & Schmitt, 1997; Lievens et al., in press; Lievens & Sackett, 2006). Finally, we believe that the SJT domain would benefit from incorporating novel presentation and response formats. Successful examples are 3D animated and avatar-based SJTs (Fetzer, 2012). Another example is the development of video-based SJTs with an open-ended response modality, namely either a written (responding in a text box) or behavioral (responding through a webcam) response format (Lievens et al., in press). Clearly, there exist a plethora of opportunities to create new SJT formats and hybrid SJTs in the future. Both practitioners and researchers should join forces to implement and examine them in the future.

References

Arthur, W., & 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

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13*, 225-232. doi: 10.1111/j.1468-2389.2005.00319.x

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance *Personnel Psychology, 62*, 229-258. doi: 10.1111/j.1744-6570.2009.01137.x

Bobko, P., Roth, P. L., & Buster, M. A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment, 13*, 1-10. doi: 10.1111/j.0965-075X.2005.00295.x

Bruce, M. M. (1974). *Examiner's manual:* Supervisory Practices Test (Rev. ed.). Larchmont, NY: Author.

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8*, 248-260. doi: 10.1111/1468-2389.00154

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

Cardall, A. J. (1942). *Preliminary manual for the Test of Practical Judgment.* Chicago: Science Research.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143-159. doi: 10.1037/0021-9010.82.1.143

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*, 233-254. doi: 10.1207/S15327043HUP1503_01

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

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410-417. doi: 10.1037/0021-9010.86.3.410

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 judgment tests. *International Journal of Selection and Assessment, 19*, 363-373. doi: 10.1111/j.1468-2389.2011.00565.x

Dean, M. A., Bobko, P., & Roth, P. L. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology, 93*, 685-691. doi: 10.1037/0021-9010-93.3.685

De Meijer, L. A. L. (2008). *Ethnicity effects in police officer selection: Applicant, assessor, and selection-method factors* (Unpublished doctoral dissertation). Erasmus University, Rotterdam.

De Meijer, L. A. L., Born, M. P., van Zielst, J., & van der Molen, H. T. (2010). Construct-Driven Development of a Video-Based Situational Judgment Test for Integrity A Study in a Multi-Ethnic Police Setting. *European Psychologist, 15*, 229-236. doi: 10.1027/1016-9040/a000027

Fetzer, M. S. (2012, April). *Current research in advanced assessment technologies*. Symposium presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

File, Q. W. (1945). The *measurement* of *supervisory quality* in *industry*. *Journal* of *Applied Psychology*, *29*, *323-337*. doi: 10.1037/h0057397

Greenberg, S. H. (1963). *Supervisory Judgment Test manual.* Washington, DC: U.S. Civil Service Commission.

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

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *International Journal of Aviation Psychology, 13*, 373-386. doi: 10.1207/S15327108IJAP1304_03

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

Lane, S., & Stone, C. A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational Measurement.* (pp. 387-431). Westport, CT: American Council on Education and Praeger.

Libbrecht, N., & Lievens, F. (2012). Validity evidence for the Situational Judgment Test paradigm to emotional intelligence measurement. *International Journal of Psychology, 47*, 438-447. doi: 10.1080/00207594.2012.682063

Libbrecht, N., Lievens, F., Carette, B., & Côté, S.C. (in press). Emotional intelligence predicts success in medical school. *Emotion*.

Lievens, F. (2006). International situational judgment tests. In J. Weekley & R. Ployhart (Eds.), *Situational judgment tests* (pp. 279-300). San Francisco: Jossey-Bass.

Lievens, F. (2013). Adjusting medical admission: Assessing interpersonal skills via situational judgment tests. *Medical Education, 47,* 182-189. doi: 10.1111/medu.12089

Lievens, F., Buyse, T., & Sackett, P. R. (2005). 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, 90*, 442-452. doi: 10.1037/0021-9010.90.3.442

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

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *Handbook of Assessment and Selection* (pp. 383-410). 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 judgment tests: a review of recent research. *Personnel Review, 37*, 426-441. doi: 10.1108/00483480810877598

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

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment 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 judgment 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 judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730-740. doi: 10.1037//0021-9010.86.4.730

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Week, J. A. (2011). Toward an Understanding of Situational Judgment Item Validity and Group Differences. *Journal of Applied Psychology, 96*, 327-336. doi: 10.1037/a0021983

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowlegde from implicit trait policies in procedural knowledge measured by a situational judgment 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 judgment 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. (2006). A theoretical basis for situational judgment tests. In J. Weekley & R. Ployhart (Eds.), *Situational judgment tests* (pp. 57-81). San Francisco: Jossey-Bass.

Northrop, L. C. (1989). *The psychometric history of selected ability constructs.* Washingtn, DC: US.S. Office of Personnel Management.

O'Connell, M. S., Hartman, N.S., McDaniel, M.A., Grubb, W.L., III, , Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment, 15,* 19-29. doi: 10.1111/j.1468-2389.2007.00364.x

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51*, 1-24. doi: 10.1111/j.1744-6570.1998.tb00714.x

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2010). Effects of individual differences on the perceived job relatedness of a cognitive ability test and a multimedia situational judgment test. *International Journal of Selection and Assessment, 18*, 394-406. doi: 10.1111/j.1468-2389.2010.00521.x

Oostrom, J. K., Born, M. Ph., Serlie, A. W., & Van der Molen, H. T. (2012). Implicit trait policies in multimedia situational judgment tests of leadership skills: Can they predict leadership behavior? *Human Performance, 25,* 335-353. doi: 10.1080/08959285.2012.703732

Oswald, F. L., Schmit, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student

performance. *Journal of Applied Psychology, 89*, 187-207. doi: 10.1037/0021-9010.89.2.187

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

Prewett, M. S., Brannick, M. T., & Peckler, B. (2013). Training teamwork in medicine: An active approach with role play and feedback. *Journal of Applied Social Psychology, 43,* 316-328. doi: 10.1111/j.1559-1816.2012.01001.x

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

Roth, P. L., Bobko, P., & Buster, M. (2013). Situational judgment 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

Roth, P. L., Bobko, P., McFarland, L., & Buster, M. (2008). Work sample tests in personnel selection: A meta-analysis of Black-White differences in overall and exercise scores. *Personnel Psychology, 61*, 637-661. doi: 10.1111/j.1744-6570.2008.00125.x

Sharma, S., Gangopadhya, M., Austin, E., & Mandal, M. K. (2013). Development and Validation of a Situational Judgment Test of Emotional Intelligence. *International Journal of Selection and Assessment, 21*, 57-73. doi: 10.1111/ijsa.12017

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

Weekley, J. A., & 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., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity,

measurement, and subgroup differences. *Human Performance, 17*, 433-461. doi: 10.1207/s15327043hup1704_5

Whetzel, D.L. & McDaniel, M.A. (2009). Situational judgment 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 judgment test performance: A meta-analysis. *Human Performance, 21*, 291-309. doi: 10.1080/08959280802137820