1-2015

Situational judgment testing: A review and some new developments

Janneke K. OOSTROM

Britt DE SOETE

Filip LIEVENS

Singapore Management University, filiplievens@smu.edu.sg

SITUATIONAL JUDGMENT TESTING:

A REVIEW AND SOME NEW DEVELOPMENTS

Janneke K. Oostrom - VU University Amsterdam

Britt De Soete - Ghent University

Filip Lievens - Ghent University

Oostrom, J. K., De Soete, B., & Lievens, F. (2015). Situational judgment testing: A review and some new developments. In I. Nikolaou, & J. K. Oostrom (Eds.), Employee recruitment, selection, and assessment: Contemporary issues for theory and practice (pp. 172-189). Sussex, UK: Psychology Press.

ABSTRACT

Although situational judgment tests (SJTs) have a long history in the personnel selection literature, there have been some recent developments in how they are designed, administered, and scored. An SJT is a measurement method typically composed of challenging work-related situations and a list of plausible courses of action. Test takers are asked to evaluate each course of action for either the likelihood that they would perform the action or the effectiveness of the action. In this book chapter, we first briefly review current practice regarding the development of SJTs in personnel selection. We also review evidence concerning reliability, construct-related validity, criterion-related validity, subgroup differences, fakability, and acceptability by test takers. Then, we focus on several promising new developments regarding the way SJTs are designed and scored. The chapter concludes with a list of areas that need to be addressed in future research.

INTRODUCTION

Situational judgment tests (SJTs) have been used for employee selection for about 80 years (e.g., McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Moss, 1926). A typical SJT presents test takers with job-related dilemmas that require relevant knowledge, skills, abilities or other characteristics to solve. The dilemmas are followed by alternative courses of action from which the test taker chooses the most appropriate response. SJTs were originally designed to sample behaviors (Motowidlo, Dunnette, & Carter 1990). Samples or simulations are based on the assumption that one can predict how well an individual will perform on the job based on a simulation of the job (McDaniel & Nguyen, 2001). As a measurement method, SJTs can be used to assess a variety of constructs (Arthur & Villado, 2008). Christian, Edwards, and Bradley (2010) showed in a review of SJT research that a

substantial number of SJTs (33%) measure heterogeneous composites. In some cases SJTs have been developed to assess specific constructs, most often leadership skills (38%) or interpersonal skills (13%).

This chapter will describe the traditional way of developing SJTs, followed by a literature review concerning how design considerations impact the quality of the SJT. Hereby we update the earlier reviews of Whetzel and McDaniel (2009) and Lievens, Peeters, and Schollaert (2008). Then, we focus on several promising new developments regarding the way SJTs are designed and scored.

SJT DEVELOPMENT

In this section, we describe current practices regarding the development of SJT items. Each item consists of a job-related dilemma, from here on named item stem, and several possible means of handling the dilemma, from here on named response options.

Development of the items

There are two popular methods for developing SJT items: critical incident and theory-based methods (Weekley, Ployhart, & Holtz, 2006). The critical incident method (Flanagan, 1954) is the most common approach used to identify the content of the items (Motowidlo, Hanson, & Crafts, 1997). The critical incidents can be collected from archival records or from interviews with subject matter experts (SMEs), for example managers, incumbents, clients, or other key stakeholders, following a format known as the antecedent-behavior-consequence (A-B-C) method (Weekley et al., 2006). The antecedents, or situational descriptors of the context leading up to the incident, are used to develop the item stem while the subsequent behavior described is used in the development of one or more of the response options. Although the critical incident approach is time-consuming and expensive, the realism of the items that are generated using this approach is likely to be high. Kanning, Grewe, Hollenberg, and Hadouch (2006) provide an example of how critical incident interviews can be used to

develop an SJT for police officers. Hunter (2003) provide an example of how archival records (i.e., a review of accident causal factors and anecdotes) can be used to develop an SJT for aviation pilots.

The second approach used to identify the content of the items is to use an underlying model (e.g., competencies identified via a job analysis, a theoretical model) and write items that reflect the dimensions of the model. If SMEs are not used to write the items, they should at least be used to review them for realism. Along these lines, Mumford, Van Iddekinge, Morgeson, and Campion (2008) provide an example of using an underlying model, in this case a team role typology, to develop an SJT measuring knowledge of team roles. Using an underlying model ensures the representativeness and job-relatedness of the SJT. However, a limitation of this approach is the lack of theory about work situations (Weekley et al., 2006).

In most cases, the items are presented by text (McDaniel & Nguyen, 2001), but it is also possible to use short videos clips (Drasgow, Olson-Buchanan, & Moberg, 1999; Weekley & Jones, 1997, 1999). Apart from the higher development costs, the use of video clips has several advantages compared with texts. First, using video clips, a richer information can be presented in the same time span because the test taker receives visual as well as auditory information (Paivio, 1986). Second, the use of video clips leads to a higher fidelity of the SJT items. The items become more realistic, making it easier for the test takers to imagine that they are actually part of the situational dilemma (Motowidlo et al., 1990). Third, the use of video clips has the advantage that test takers are not required to read lengthy texts (Chan & Schmitt, 1997).

Response instructions

After developing the SJT items, the response instructions have to be determined. There are two types of response instructions that can be used: knowledge-based and behavioral tendency instructions (McDaniel & Nguyen, 2001). Knowledge-based response instructions,

also known as ‘should-do’ response instructions, ask the test taker to identify the best or correct course of action in the given situation. Behavioral tendency response instructions, also known as ‘would-do’ response instructions, ask the test taker to express how he or she would likely behave in the given situation (McDaniel, Hartman, & Whetzel, & Grubb, 2007). The two instruction types relate to the distinction between typical and maximal performance (Cronbach, 1984). Maximal performance tests assess test takers’ performance when doing their best and are generally used to make inferences about ability. Typical performance tests assess how test takers typically behave and are generally used to make inferences about personality, attitudes, and other non-cognitive aspects. SJTs with knowledge response instructions are maximal performance tests as test takers make judgments about what constitutes effective performance. SJTs with behavioral tendency response instructions are typical performance tests as test takers report how they typically behave (McDaniel et al., 2007).

Scoring methods

A final aspect to consider when developing SJTs, is how to score test takers’ answers. At least three different methods for determining the effectiveness of the response options have been explored in the literature, i.e., expert-based, empirical-based, and theory-based methods. Note that it is also possible to combine some of these methods. In that case, a hybrid scoring method is used.

The most common scoring approach in the SJT literature is asking SMEs to make judgments concerning the effectiveness of the response options (e.g., Lievens et al., 2008; McDaniel & Nguyen, 2001). These judgments are pooled subsequently either using consensus or actuarial methods (McDaniel & Nguyen, 2001). Although the results with the expert-based scoring method are generally positive (e.g., McDaniel et al., 2007; Krokos, Meade, Cantwell, Pond, & Wilson, 2004), this approach has several drawbacks (Lievens, 2000). When SJTs are

expert-based scored, the test taker's score represents the level of agreement with the judgments of the SMEs, and therefore is dependent on the unique perspectives of the SME group (Krokos et al., 2004). It is likely that different groups of SMEs derive different keys. A final drawback is that it can be difficult to gain agreement among SMEs regarding the effectiveness of the response alternatives (McHenry & Schmitt, 1994).

There are two different empirical-based scoring methods, namely external and internal. When SJTs are externally scored, they usually are administered to a large pilot sample (Lievens et al., 2008). Based on the correlation with a criterion measure, items are selected and weighted. The crucial issue in external scoring is the quality of the criterion. If the criterion is deficient, contaminated, or biased, empirical keys will reflect these problems in the scoring structure (Mumford & Owens, 1987). External scoring approaches are rarely used for SJTs. Dalessio (1994) presents one of the few examples of an empirical scoring method for an SJT to predict turnover among insurance agents. The internal approach requires test items being scored in terms of their interrelationships. Factor analytic procedures are used to create subscales which may then be combined for prediction in a multiple regression (Schoenfeldt & Mendoza, 1994). One of the advantages of this scoring approach is that the items can be scored and weighted taking account of their relationship with the other items and that the number of items can be reduced. A drawback is that the factors may be difficult to interpret, especially when heterogeneous item pools are used (Lievens, 2000). We were able to trace only one study on SJTs in which an internal scoring approach is used, namely the study of Lievens (2000), who developed and applied an empirically based scoring procedure based on a multiple correspondence analysis on an SJT for sales performance. Although empirical-based scoring methods often have high validity (e.g., Bergman, Drasgow, Donovan, Henning, & Juraska, 2006), the method is criticized for being atheoretical. Furthermore, the

method is questioned regarding its generalizability and stability (Mumford & Owens, 1987), and capitalization on chance (Bergman et al., 2006).

The third and last frequently used method of developing scoring keys is to rely on an underlying model. This scoring method is often used when the response options are already constructed to reflect a theoretical model. Bergman et al. (2006) describe an SJT in which the response options reflect three graduated levels of delegation of decision-making to the team and used Vroom's contingency model to score test takers' answers (Vroom & Jago, 1978). Theory-based scoring methods are more likely to generalize. Yet, the crucial issue in external scoring is the quality of the theory, which might be flawed or fundamentally incorrect (Bergman et al., 2006).

SJT CHARACTERISTICS AND THEIR IMPACT ON SELECTION TEST CRITERIA

As described above, many choices have to be made when developing SJTs. It is important to know how these design considerations impact the quality of the SJT as a tool in selecting new employees. In this section, we describe how these design considerations affect six important selection test criteria.

Reliability

Regarding SJTs, the most widely used measure of reliability is the internal consistency reliability as indexed by coefficient alpha. However, estimating the internal consistency of SJT scores is often problematic and not very relevant, because most SJTs – specifically those SJTs that are developed using the critical incident method – tend to assess multiple constructs (McDaniel & Whetzel, 2005). As a result, over the years many researchers have suggested that test-retest reliability is a better estimate of SJT score reliability (e.g., McDaniel et al., 2007; Motowidlo et al., 1990).

Ployhart, Campion, and MacKenzie (2012) have conducted a meta-analysis on SJT reliability coefficients and found a mean test-retest reliability of .61. However, they were able

to trace only eight studies in which the rest-retest reliability coefficient was mentioned. Ployhart and Ehrhart (2003), who compared one SJT with six different response instructions, found significant differences in test-retest reliability coefficients; behavioral tendency response instructions showed higher test-retest reliabilities than knowledge-based response instructions. However, these results should be interpreted with caution as the analyses were based on small samples ranging from 21 to 30.

SJTs that are developed based on an underlying theory are expected to show higher internal consistency, as the items are more likely to load highly on one or more factors (Ployhart et al., 2012). Yet, no systematic research exists wherein development procedures or different scoring methods are compared in terms of reliability.

Construct-related validity

The construct-related validity of SJTs remains hard to pin down. According to Stemler and Sternberg (2006) SJTs measure practical intelligence, which is the ability to adapt to, shape, and select everyday environments. However, most researchers argue that SJT performance can be determined by a number of constructs such as cognitive ability, personality, and job experience (Weekley & Jones, 1999). For SJTs that are developed based on an underlying theory, it should evidently be clearer which constructs they are measuring. However, most SJTs in which the item stems and/or response options reflect different dimensions failed to provide reliable subscores reflecting these dimensions (e.g., Weekley et al., 2006).

Almost all construct-related validity evidence until now has been restricted to paper-and-pencil SJTs. The test medium is expected to affect the construct-related validity (McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006). For example, video-based SJTs are expected to reduce the cognitive load of an SJT primarily by reducing the reading demands. Chan and Schmitt (1997) demonstrated that reading comprehension correlated positively with

performance on a paper-and-pencil SJT, but was nearly uncorrelated with performance on a video-based version of the same SJT. Similarly, Lievens and Sackett (2006) found that cognitive ability correlated positively with performance on a paper-and-pencil SJT but not with performance on a video-based version of the same SJT.

The response instruction has also been found to affect the SJT's construct validity. The meta-analysis of McDaniel et al. (2007) showed that SJT scores with knowledge-based response instructions correlate more highly with cognitive ability scores than SJTs with behavioral tendency response instructions, whereas SJT scores with behavioral tendency response instructions correlate more highly with personality ratings. This is in line with the notion that SJTs with knowledge-based response instructions tap more into maximal performance and SJTs with behavioral tendency response instructions tap more into typical performance (McDaniel et al., 2007). Test developers should, therefore, choose the type of instructions on the basis of the type of performance they wish to emphasize in their assessment (Whetzel & McDaniel, 2009).

Criterion-related validity

In general, the literature has found SJT scores to have good predictive validities (e.g., Christian et al., 2010). McDaniel et al. (2007) demonstrated in their meta-analysis that SJT scores have an average observed validity of .20, and have incremental validity over cognitive ability scores and Big Five personality ratings. There is no systematic research in which the design procedures (critical incident and theory-based methods) are compared. Yet, the effects of the other design features on SJT criterion-related validity have been examined.

Christian et al. (2010) meta-analytically showed that video-based SJTs have higher validity than paper-and-pencil SJTs for predicting interpersonal skills. That is, video-based SJT scores of interpersonal skills had an average validity of .47, which was significantly higher than the average validity of .27 for paper-and-pencil SJT scores of interpersonal skills.

The meta-analysis of McDaniel et al. (2007) showed that response instructions had little moderating effect on criterion-related validity. Note that most studies included in these meta-analyses are based on incumbent samples. More recently, Lievens, Sackett, and Buyse (2009) conducted a study on the moderating effect of response instructions on criterion-related validity in a large-scale high-stakes selection context. Their results corroborated the findings of McDaniel et al.; no moderating effect of response instructions on criterion-related validity was found.

Several studies have shown that empirical-based scoring methods and expert-based scoring methods have similar levels of validity (e.g., Bergman et al., 2006; MacLane, Barton, Holloway-Lundy, & Nickles, 2001; Weekley & Jones, 1999). Criterion-related validity results regarding the theory-based scoring method are inconsistent (e.g., Bergman et al., 2006; Olson-Buchanan et al., 1998). Clearly more research is needed to better understand when theory-based scoring methods work best.

McDaniel, Psotka, Legree, Yost, and Weekley (2011) describe two adjustments to common scoring approaches which improve the criterion-related validity of the SJT. The first adjustment - which is only applicable to SJTs that use Likert scales - is to standardize scores using a within-person *z* transformation, so that all test takers have the same mean and *SD* across items. This transformation removes information related to elevation (i.e., the mean of the items for a test taker) and scatter (i.e., the magnitude of a test taker's score deviations from his or her own mean). Elevation and scatter are a source of systematic error as they often reflect response tendencies, such as a preference for using extreme ends of the scale. McDaniel et al. (2011) demonstrated that controlling for elevation and scatter resulted in substantial improvements to item validity. The second adjustment is to drop response options with midrange means because these response options tend to provide little information on

whether the test taker is able to identify (in)effective behavior. McDaniel et al. showed that dropping midrange items permits the SJT to be shortened without harming validity.

Ethnic score differences

SJTs appear to display smaller ethnic score differences than cognitive ability tests, which makes them an attractive selection tool. Whetzel, McDaniel, and Nguyen (2008) reported in their meta-analysis a Black-White score difference of 0.38 *SD* and a Hispanic-White score difference of 0.24 *SD*, in favor of Whites. Research on ethnic SJT score differences in Europe revealed comparable findings, with ethnic minorities obtaining systematically somewhat lower scores than majority test takers (d = 0.38; De Meijer, 2008).

Research on ethnic score differences on selection tools has repeatedly shown that the instrument's cognitive loading constitutes one of the most important drivers of ethnic score differences (e.g., Bobko, Roth, & Buster, 2005; Dean, Bobko, & Roth, 2008). In this context, SJTs with a higher cognitive loading have been found to display larger ethnic score differences than SJTs with a lower cognitive loading (Roth, Bobko, & Buster, 2013; Whetzel et al., 2008). A promising strategy to reduce ethnic score difference on SJTs is by using video-based items instead of paper-and-pencil items, as this results in lower reading demands and therefore a lower cognitive loading (Chan & Schmitt, 1997; Lievens & Sackett, 2006). Along these lines, Chan and Schmitt (1997) found that video-based SJTs displayed significantly smaller ethnic score differences than content-wise identical paper-and-pencil SJTs (d = 0.21 versus d = 0.95). Personality loading has also been found to influence the magnitude of ethnic score differences. Black-White score differences demonstrated to be larger when the SJT is characterized by a lower emotional stability loading, whereas Hispanic-White score differences tend to increase with lower agreeableness and conscientiousness loadings (Whetzel et al., 2008).

The type of response instructions has been found to influence the size of ethnic score differences (Nguyen & McDaniel, 2003; Whetzel et al., 2008). Whetzel et al. (2008) showed that SJTs with knowledge-based instructions consistently display larger differences than SJTs with behavioral tendency instructions for Black-White, Hispanic-White and Asian-White score comparisons. This finding can in most cases be attributed to the larger cognitive loading of knowledge-based response instructions (Nguyen & McDaniel, 2003).

Finally, the scoring method has proven to influence ethnic score differences. As mentioned above, to increase the criterion-related validity of SJTs with Likert scales, McDaniel et al. (2011) suggested to control for elevation and scatter by using a within-person *z* transformation. An additional benefit of this adjustment is that score differences arising as a result of Black-White discrepancies in extreme responding are reduced. In a first study, Black-White ethnic score differences decreased from $d = 0.43$ to $d = 0.29$. A second study yielded similar results with *d* decreasing from 0.56 to 0.36.

The effect of the development procedure, more specifically the influence of the cultural (dis)similarity of the SMEs involved in SJT developing and scoring, on the magnitude of ethnic score differences is still unknown. Additionally, as most studies on ethnic score differences are performed in a U.S. context, systematic research incorporating other ethnic minority groups than Blacks and Hispanics is rather limited.

Fakability

Faking on a selection test can be defined as an applicants' conscious distortion of their answers to score more favorably (e.g., McFarland & Ryan, 2000). Although there is an ongoing debate on whether faking influences a selection test's criterion-related validity (e.g., Hough, 1998; Ones & Viswesvaran, 1998), researchers do agree that faking can have a significant effect on who is hired.

As far as we know, there are no studies on the influence of the development procedure or scoring method of the SJT on its fakability. Nevertheless, it seems plausible that the constructs measured, the development of response options, and the scoring method affect an SJT's fakability. SJTs that tap into less fakable domains such as cognitive ability should be less susceptible to faking than those that tap into domains such as personality (Hooper, Cullen, & Sackett, 2006). When the response options reflect dimensions of an underlying model and the model is used to score test takers' answers, the SJT is expected to be more susceptible to faking due to its greater transparency (Hough & Paullin, 1994). Weekley et al. (2006) argue that test developers should be able to control the SJT's fakability by developing and selecting response options with comparable social desirability, so that test takers are not easily able to identify the correct response.

McDaniel et al. (2011) showed that standardizing SJT scores using a within-person *z* transformation - which is only applicable to SJTs that use Likert scales - reduces the coachability of SJTs. Like faking, coaching may lead to the hiring of individuals whose true score is less than what it appears to be. McDaniel et al. found that the coaching strategy of avoiding extreme responses, which is generally an effective strategy (Cullen, Sackett, & Lievens, 2006), is ineffective for the standardized scales and even lowered scores up to 1.07 *SD*.

A few studies have been conducted regarding the effects of response instruction on the SJT's fakability. Nguyen, Biderman, and McDaniel (2005) found that test takers could distort their answers on an SJT with behavioral tendency instructions such that on average they were able to elevate their scores with 0.15 or 0.34 *SD*, depending on whether they took the SJT in the honest or faking condition first. As it is difficult to fake knowledge, the results for the SJT with knowledge instructions were inconsistent; faking even lead to lower scores when test takers had to answer honestly first. Peeters and Lievens (2005) conducted a between-subjects

study on the fakability of SJTs with behavioral tendency instructions and found that the test takers in the fake condition scored 0.89 *SD* higher than the test takers in the honest condition. Furthermore, they found that faking had a negative effect on the criterion-related validity of the SJT. Note that these effect sizes are derived from experimental faking research. The effect sizes are likely to be different in an applicant sample. Lievens et al. (2009) found that in such a context test takers respond similarly to an SJT with behavioral tendency instructions and an SJT with knowledge-based instructions.

Test taker perceptions

Previous studies have demonstrated that test takers' perceptions are related to numerous outcomes, such as intentions to accept the job, the likelihood of litigation against the outcome of the selection procedure, and perceived organizational attractiveness (e.g., Anderson, Lievens, Van Dam, & Ryan, 2004; Ryan & Ployhart, 2000). Systematic research on the effects of the development procedure, response instructions, and scoring method on test taker perceptions is lacking. However, a fair amount of research has been conducted on the effects of stimulus format on test taker perceptions. Video-based SJTs provide a realistic job preview and therefore are expected to be more attractive for test takers in terms of interest and motivation than paper-and-pencil SJTs. Richman-Hirsch, Olson-Buchanan, and Drasgow (2000) demonstrated that compared to a paper-and-pencil SJT, the video-based version with identical content indeed yielded more positive reactions. The video-based SJT was perceived as more content valid, more face valid, more enjoyable, and led to more satisfaction with the assessment process. Chan and Schmitt (1997) demonstrated that test takers rate the face validity of a video-based SJT significantly more positively than the face validity of a paper-and-pencil SJT. Kanning et al. (2006) examined reactions to SJT items that differed with regard to interactivity (non-interactive versus interactive) and medium (video versus paper-and-pencil). Video-based SJT items, in which the response of the participants determines the

further course of the item, were perceived as the most favorable in terms of enjoyment, acceptance, and job relatedness.

Table 1 presents an overview of the research findings regarding the impact of design characteristics on the six key criteria for selection tests. As has become apparent, there are many gaps in the literature. More systematic research is needed to establish consensus regarding optimal SJT development methods.

Table 1.

Impact Design Characteristics on Selection Test Criteria

Selection test criterion	**Development method**	**Response instructions**	**Scoring method**	**Key references**
Reliability	Unknown	Some evidence for higher test-retest reliability behavioral tendency instructions	Unknown	Ployhart et al. (2012), Ployhart & Ehrhart (2003)
Construct-related validity	Video-based SJTs have lower cognitive loading than paper-and-pencil SJTs	Knowledge-based instructions capture maximal performance and behavioral tendency instructions capture typical performance	Unknown	Chan & Schmitt, (1997), Lievens & Sackett (2006), McDaniel et al. (2007)
Criterion-related validity	Video-based SJTs have higher validity for interpersonal skills than paper-and-pencil STJs	No moderating effects	Some evidence that empirical-based methods and expert-based methods have higher validity than theory-based methods	Bergman et al. (2006), Christian et al. (2010), Lievens et al. (2009), McDaniel et al. (2007), McDaniel et al. (2012)
			Scoring adjustments (within-person z transformation and removing items with midrange means) lead to higher validity	
Ethnic score differences	Video-based SJTs show smaller ethnic score differences than paper-and-pencil SJTs	Behavioral tendency instructions lead to lower ethnic score differences than knowledge-based instructions	Scoring adjustments (within-person z transformation) lead to smaller ethnic score differences	Chan & Schmitt (1997), McDaniel et al. (2011), Whetzel et al. (2008)
Fakability	Unknown	Knowledge-based instructions are less	Some evidence that theory-based	Hough & Paullin (1994), McDaniel et

		fakable than behavioral tendency instructions	methods lead to higher susceptibility to faking than other scoring methods Scoring adjustments (removing items with midrange means) reduces coachability	al. (2011), Ngyen et al. (2005), Peeters & Lievens (2005)
Acceptability	Video-based and interactive SJTs lead to more positive test taker perceptions	Unknown	Unknown	Chan & Schmitt (1997), Kanning et al. (2006), Richman-Hirsch et al. (2000)

NEW DEVELOPMENTS

Recently, there have been new developments in the way SJTs are developed and scored. In this section, we describe three important advancements that aim at improving the construct and criterion-related validity of SJTs.

A construct-based approach

Based on their meta-analysis, Christian et al. (2010) argue that SJT research could benefit from a construct-based approach. So far, there has been a lack of attention to SJT constructs (Arthur & Villado, 2008; Schmitt & Chan, 2006). Many studies fail to report the constructs measured by SJTs (e.g., Cucina, Vasipoulos, & Leaman, 2003; Pereira & Schmidt, 1999) and even when SJTs are developed to assess one or more specific constructs overall scores rather than scores for specific constructs are reported (e.g., Chan & Schmitt, 2002; Weekley & Jones, 1997, 1999). A construct-based approach offers several theoretical and practical advantages: 1) the specification of the construct domain helps to reduce contamination due to the measurement of unintended, non-job-relevant constructs (Christian et al., 2010), 2) the items of the SJT will load highly on one (or more) factors and exhibit little item-specific variance SJTs leading to higher reliability coefficients (Ployhart et al., 2012), 3) it provides insight into why the SJT is related to the criterion of interest (Arthur & Villado, 2008; Schmitt & Chan, 1998), and 4) it provides the opportunity to conceptually match the predictor and criterion domain (Paunonen, Rothstein, & Jackson, 1999).

De Meijer, Born, Van Zielst, and Van der Molen (2010) developed an SJT to measure the construct of integrity and Bledow and Frese (2009) developed an SJT to measure the construct of personal initiative. Both found support for the convergent and divergent validity of SJT scores. Furthermore, De Meijer et al. (2010) report an internal consistency coefficient of .69. These results demonstrate that it is possible to develop an SJT that assesses a specific construct. However, not all attempts have been successful (e.g., Pulakos & Schmitt, 1996). According to Ployhart, Porr, and Ryan (2004), this is because most recent studies have used minor variations of the method of developing SJT items described above. Ployhart et al. (2004) describe an alternative way of developing SJTs to assess specific constructs. The steps are: 1) defining the performance domain and indentifying relevant criterion behaviors, 2) identifying situations that results in the maximal variability in behaviors such that the trait(s) of interest can be manifested, 3) linking the situations to the criterion behaviors, 4) constructing response options that lie on a continuum with each response option reflecting a different level of the trait, and 5) asking experts to rate the situations and the response options for their relevance to the trait(s) of interest. Ployhart et al. used this approach to develop and SJT for neuroticism, agreeableness, and conscientiousness. Their results suggested that SJT items can be written to reflect personality traits and that such an SJT shows adequate criterion-related validity.

The use of alternative response formats

There are two recent developments regarding the response format of SJTs. The first development aims at increasing the fidelity of the SJT by using a constructed response format instead of a multiple choice format. Although a multiple choice format has several advantages over a constructed response format such as the possibility to administer the test to large groups at the same time and the cost-effectiveness in scoring test takers' answers (Edwards, Arthur, & Bruce, 2012; Motowidlo et al., 1990), the format does not correspond with real-life.

In addition, a multiple choice format is susceptible to guessing and other test-taking strategies (Ellis & Ryan, 2003). In so-called constructed response SJTs, challenging job-related scenarios are presented by using video clips. After the scenario is presented, applicants are asked to act out their response, while being filmed by a webcam (Oostrom, Born, Serlie, & Van der Molen, 2010). Although such a format is less standardized and therefore more expensive and time-consuming to score as compared to a multiple choice format, it invokes greater realism and fidelity than a multiple choice response format. Subsequently, it is typically perceived more positively by test takers. Particularly ethnic minority test takers, who might have negative experiences with multiple choice tests, seem to appreciate tests with constructed response formats (Edwards & Arthur, 2007; Ryan & Ployhart, 2000). Furthermore, constructed response SJTs have been found to be predictive of various criteria such as employment agents' job placement success (Oostrom et al., 2010) and learning activities of students (Oostrom, Born, Serlie, & Van der Molen, 2011), training performance ratings of policemen (Lievens, De Corte, & Westerveld, in press), and contextual job performance ratings of government employees (De Soete, Lievens, & Oostrom, 2013). Effects on ethnic score differences have been promising, with constructed response SJTs displaying ethnic score differences of 0.14 *SD* (De Soete, Lievens, Oostrom, & Westerveld, in press).

The second development regarding the response format of SJTs is presenting one response option instead of multiple, usually 3 to 12, response options per item. Motowidlo and colleagues (Crook et al., 2011; Martin & Motowidlo, 2010; Motowidlo, Crook, Kell, & Naemi, 2009) have developed several of these so-called single response SJTs. They argue that the development and scoring of single response SJTs is less labor intensive than the development of traditional SJTs as it eliminates the need for SMEs to generate behavioral responses to situations and minimizes the time needed to rate multiple response options for effectiveness. Moreover, with single response SJTs the items can be more easily classified to

a criterion dimension, which is likely to improve the construct-related validity of the SJT and allows for a better predictor-criterion alignment. Initial evidence is promising. Motowidlo et al. (2009) showed that a single response SJTs is able to predict the work effort of volunteers. Crook et al. (2011) showed that a single response SJTs is a valid predictor of tour guide performance at a children's museum.

Implicit trait policies

To explain why SJTs are often correlated with measures of personality traits, Motowidlo, Hooper, and Jackson (2006) developed the implicit trait policy (ITP) theory. ITPs are the implicit beliefs of individuals about the effectiveness of different levels of trait expression. For instance, an individual may believe that the expression of conscientiousness is generally very effective. ITPs are measured by correlating test takers' effectiveness ratings of SJT response options with the level of trait expression of these response options. The central proposition of the ITP theory is that individual differences in personality traits affect judgments of the effectiveness of SJT response options that express those personality traits. Motowidlo et al. (2006) found empirical support for their theory, as they were able to demonstrate that ITPs for agreeableness, conscientiousness, and extraversion are related to individual differences in these personality traits. Furthermore, Motowidlo and Beier (2010) demonstrated that ITPs are able to predict a performance composite based on supervisor ratings. Similarly, Oostrom, Born, Serlie, and Van der Molen (2012) demonstrated that an SJT for leadership skills can be used to measure individual differences in ITPs and that those ITPs are able to predict leadership behavior over and above leadership experience and personality traits.

The ITP theory also provides practitioners an alternative scoring method for SJTs, by which this general domain knowledge about the costs and benefits of expressing particular personality traits can be measured. There are several advantages of using this alternative

scoring method. First, scoring keys for ITPs do not require experts with considerable domain-specific knowledge and experience. Second, as ITPs tap general domain knowledge, the validity of ITPs for targeted traits may be more generalizable across job domains than the validity of traditionally scored SJTs.

SUGGESTIONS FOR FUTURE RESEARCH

From our review of the literature on the development and scoring of SJTs, it has become clear that there are several pressing research needs. First of all, much more systematic studies are needed in which the different development methods, response instructions, and scoring methods are compared in terms of reliability, validity, ethnic score differences, and test taker reactions. Consensus regarding optimal SJT development methods is a prerequisite to establishing SJTs as a mean to measure and predict specific constructs. These studies should consider using a construct-based approach. A construct-based approach offers several theoretical and practical advantages such as the ability to generalize findings across time and jobs (Arthur & Villado, 2008; Schmitt & Chan, 1998).

We also presented several new developments which we believe will help improve SJTs. Yet, more research on these trends is welcomed. Ployhart et al. (2004) have presented an alternative way of developing construct-based SJTs and Motowidlo et al. (2006) have presented an alternative scoring method for SJTs by which ITPs can be measured. Although researchers have called for a more construct-based approach in SJT research (e.g., Christian et al., 2010), these alternative development and scoring methods are not yet widespread. Studies are needed to compare the usability of alternative development and scoring methods to that of traditional methods. Future studies should also look into the boundary conditions of these alternative methods. For example, it might be that the alternative SJT development method of Ployhart et al. is more suited for the assessment of constructs that lie on a continuum, such as personality, than for other constructs. In addition, it might make the SJT more fakable.

Two promising alternative response formats have been presented, that is the use of constructed response formats and single-response formats. Future studies should compare constructed response SJTs to traditional multiple-choice SJTs in terms of validity, ethnic score differences, and test taker perceptions. Motowidlo and colleagues (Crook et al., 2011; Martin & Motowidlo, 2010; Motowidlo et al., 2009) have developed so-called single response SJTs which are less labor intensive to develop than traditional SJTs. So far, results have been promising, which should encourage future studies on the development of single response SJTs.

CONCLUSION

In this chapter, we have reviewed the traditional way of developing and scoring SJTs and how different development and scoring procedures affect the SJT's reliability, validity, ethnic score differences, fakability, and acceptability. Clearly, more systematic research is needed in which the different development and scoring procedures are compared. Consensus regarding optimal SJT development methods is important to establish SJTs as a mean to measure and predict specific constructs. We also presented several new developments, namely the use of a construct-based approach, constructed response formats, single-response formats, and ITPs. We believe these developments will help improve SJTs. Yet, more research-based evidence is needed to evaluate their viability.

REFERENCES

Anderson, N. R., Lievens, F., van Dam, K., & Ryan, A. M. (2004). Future perspectives on employee selection: Key directions for future research and practice. *Applied Psychology: An International Review*, *53*, 487-501.

Arthur, W. J., & Villado, A. (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.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*, 223-235.

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, *62*, 229-258.

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.

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.

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

Christian, M. S., Edwards, J. C., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83-117.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York, NY: Harper & Row.

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.

Cucina, J. M., Vasilopoulos, N. L., Leaman, J. A. (2003, April). *The bandwidth-fidelity dilemma and situational judgment test validity*. Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14*, 142-155.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology*, *9*, 23-32.

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.

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. Ph., Van Zielst, J. & Van der Molen, H.T. (2010). The construct-driven development of a video-based situational judgment test measuring integrity: A study in a multi-ethnic police setting. *European Psychologist, 15*, 229-236.

De Soete, B., Lievens, F., & Oostrom, J. K. (April, 2013). *The diversity-validity dilemma in selection: The role of response fidelity.* Poster presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, Texas, Houston.

De Soete, B., Lievens, F., Oostrom, J. K., & Westerveld, L. (in press). Alternative predictors for dealing with the diversity-validity dilemma in personnel selection: The constructed response multimedia test. *International Journal of Selection and Assessment*.

Drasgow, F., Olson-Buchanan, J. B., & Moberg, P. J. (1999). Development of interactive video assessments. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 177-196). Mahwah, NJ: Erlbaum.

Edwards, B. D., & Arthur, W. J. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology*, *92*, 794-801.

Edwards, B. D., Arthur, W., & Bruce, L. L. (2012). The three-option format for knowledge and ability multiple-choice tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment, 20*, 65-81.

Ellis, A. P. J., & Ryan, A. M. (2003). Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-efficacy. *Journal of Applied Social Psychology, 33*, 2607-2629.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance*, *11*, 209-244.

Hough, L. M, & Paullin, C. (1994). Construct-oriented scale construction: The rational approach. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 109-145). Palo Alto, CA: CPP Books.

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

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

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment*, *2*3, 168-176.

Krokos, K. J., Meade, A. W., Cantwell, A. R., Pond, S. B., & Wilson, M. A. (2004, April). *Empirical keying of situational judgment tests: Rationale and some examples*. Paper presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Lievens, F. (2000). Development of an empirical scoring scheme for situational inventories. *European Review of Applied Psychology*, *50*, 117-124.

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*.

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. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91*, 1181-1188.

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

MacLane, C. N., Barton, M. G., Holloway-Lundy, A. E., & Nickles, B. J. (2001, April). *Keeping score: Expert weights on situational judgment responses.* Paper presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Martin, M. P., & Motowidlo, S. J. (2010, April). A single-response SJT for measuring procedural knowledge for human factors professionals. Paper presented at the 25th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (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.

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

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96*, 327-336.

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence*, *33*, 515-525.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N. T., & Grubb, W. L., III. (2006). Situational judgment tests: Validity and an integrative model. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 183-204). Mahwah, NJ: Lawrence Erlbaum.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*, 812-821.

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 193-232). Hillsdale, NJ: Lawrence Erlbaum.

Moss, F. A. (1926). Do you know how to get along with people?. *Scientific American*, *135*, 26-27.

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

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.

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.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 241-260). Palo Alto, CA, Davies-Black Publishing.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*, 749-761.

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement*, *11*, 1-31.

Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The Team Role Test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology*, *93*, 250-267.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment*, *13*, 250-260.

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

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.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, *11*, 245-269.

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

Oostrom, J. K., Born, M. Ph., Serlie, A. W., & Van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology*, *10*, 78-88.

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

Paivio, A. (1986). *Mental representation: A dual coding approach*. Oxford, UK: University Press.

Paunonen, S. V., Rothstein, M. G., & Jackson, D. N. (1999). Narrow reasoning about the use of broad personality measures for personnel selection. *Journal of Organizational Behavior*, *20*, 389-405.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*, 70-89.

Pereira, G. M., & Schmidt, H. V. (1999, April). *Situational judgment tests: Do they measure ability, personality, or both.* Paper presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Ployhart, R. E., Campion, M. C., & MacKenzie, W. I. (2012, April). *Reliability and situational judgment tests: A review of the literature.* Paper presented at the 27th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

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.

Ployhart, R. E., Porr, W. B., & Ryan, A. M. (2004). *A construct-oriented approach for developing situational judgment tests in a service context*. Unpublished manuscript.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance*, *9*, 241-258.

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.

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.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management*, *26*, 565-606.

Schmitt, N., & Chan, D. (Eds.) (1998). *Personnel selection: A theoretical approach*. Thousand Oakes, CA: Sage.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135-155). Mahwah, NJ: Lawrence Erlbaum.

Schoenfeldt, L. F., & Mendoza, J. L. (1994). Developing and using factorially derived biographical scales. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 147-169). Palo Alto, CA: CPP Books.

Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 107-131). Mahwah, NJ: Lawrence Erlbaum.

Vroom, V. H., & Jago, A. G. (1978). On the validity of the Vroom-Yetton model. *Journal of Applied Psychology*, *63*, 151-162.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25-49.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology*, *52*, 679-700.

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

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

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.