Situational Judgment Tests:

From Low-fidelity Simulations to Alternative Measures of Personality and the Person-Situation Interplay

Filip Lievens, Philipp Schäpers & Christoph N. Herde
Singapore Management University


This chapter is not the copy of record and may not exactly replicate the final, authoritative version of the chapter.

Address for correspondence: Prof. Dr. Filip Lievens, Singapore Management University, Lee Kong Chian School of Business, 50 Stamford Road, 178899, Singapore, filiplievens@smu.edu.sg
Abstract

In employment and education settings, Situational Judgment Tests (SJT) have made strong inroads. So far, however, they are still underutilized in personality research. The objective of this chapter is to outline how SJTs might be adapted to measure personality traits, shed light onto the person-situation interplay, and stimulate research on it. We start by discussing the traditional simulation-based view on SJTs, including information on their development and research results. Next, we show how more recent versions have started to assess people’s knowledge of relevant behavior related to personality traits. Finally, we specify various strategies as to how SJTs might be further adapted to shed light on the personality-situation interplay. Along these lines, we show how SJTs might be used to assess within-person variability across situations, situation-trait contingencies, proactive transactions, behavioral responses, narratives and goals, and personality disorders.

**Keywords**: Situational Judgment Tests, personality-situation interplay, simulation, implicit trait policies
Bio of the authors

Filip Lievens is Lee Kong Chian Professor of Human Resources at the Lee Kong Chian School of Business of Singapore Management University. He is also an Honorary Professor at the University of Cape Town and a visiting professor at Ghent University, Belgium. In 1999, he obtained his PhD at Ghent University, Belgium.

His main interests deal with talent acquisition, talent assessment, and adaptability. As an internationally recognized scholar, he has published over 45 papers in top-tier journals, including the Annual Review of Psychology, Journal of Applied Psychology, Personnel Psychology, Journal of Management, Organizational Behavior and Human Decision Processes, Intelligence, and Organizational Research Methods.

Filip Lievens serves on the editorial board of Journal of Applied Psychology and Personnel Psychology. He was the first European winner of the Distinguished Early Career Award of the Society for Industrial and Organizational Psychology. He is also a winner of the Jeanneret Award and the Douglas Bray and Ann Howard Award, which recognize outstanding contributions in the assessment domain. He was a recipient of the Friedrich Wilhelm Bessel-Forschungspreis of the Alexander von Humboldt- Foundation and is laureate of the Royal Flemish Academy of Sciences and Arts, an award to value substantial contributions in any discipline in social sciences. He is a fellow of the Society for Industrial and Organizational Psychology (SIOP), the American Psychological Association (APA), the Association for Psychological Science (APS), and the International Association of Applied Psychology (IAAP).

Philipp Schäpers is currently working as a Postdoctoral Research Fellow at the Lee Kong Chian School of Business of Singapore Management University. His research is situated in the field of personnel selection and psychological assessment. He is interested in the underlying working mechanisms of selection and testing procedures and his work seeks to
improve the theoretical and practical understanding of selection tools. For instance, his PhD thesis revolved around the question of how situations in Situational Judgment Tests affect their psychometric properties. He has published in several scientific journals, including *Journal of Applied Psychology, Personnel Psychology, and Journal of Occupational and Organizational Psychology*. Further interests are diversity management, digitalization, and employer branding.

**Christoph N. Herde** is currently working as a Research Fellow at the Lee Kong Chian School of Business of Singapore Management University. His main interests deal with personnel selection with special attention to modular approaches of selection methods and simulation-based methods, such as Multiple Speed Assessments, Situational Judgment Tests, or Assessment Centers. Further interests are within-person variability in behavior and performance as well as how this variability can be captured and used to assess constructs like (interpersonal) adaptability, learning agility, or stress resilience. Other research interests are research methods and statistical modeling.
Introduction

Situational Judgment Tests (SJT) are popular instruments in personnel selection that assess an applicant's judgment regarding critical work-related situations (e.g., Weekley, Ployhart, & Holtz, 2006; Whetzel & McDaniel, 2009). SJTs confront applicants with short job-related situations and different response options to indicate how they would react to these incidents (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Reasons for SJTs’ popularity are manifold: for instance, SJTs show substantial validity for the prediction of job performance (e.g., Christian, Edwards, & Bradley, 2010; McDaniel, Hartman, Whetzel, & Grubb III, 2007), favorable applicant perceptions (Kanning, Grewe, Hollenberg, & Hadouch, 2006), are less fakeable than self-reports (Hooper, Cullen, & Sackett, 2006; Kasten, Freund, & Staufenbiel, 2018), and are less costly than high-fidelity simulations (e.g., assessment-centre exercises; Lievens & Patterson, 2011). Notably, SJTs have also emerged as a valid tool for assessment and selection in medical and educational settings (e.g., admission and credentialing purposes).

The growing interest in SJT research is also attested to by an increasing number of papers and citations related to SJTs. Since Motowidlo, Dunnette, and Carter (1990) reinvigorated interest in SJTs, researchers have published 182 papers about SJTs. These papers have been cited 2,854 times in 1,702 articles as of December 2018. As shown in Figure 1, research interest in SJTs has constantly increased as indicated by researchers conducting and citing a growing number of SJT publications.

INSERT FIGURE 1 HERE

In personality research, scenario-based approaches like SJTs are less well known. Recently, there has been growing interest in adopting SJTs for personality research (e.g., Lievens, 2017a; Mussel, Gatzka, & Hewig, 2018; Oostrom, de Vries, & de Wit, 2019). These researchers argued that personality research can benefit from scenario-based measures of
personality, such as SJTs, in order to complement common approaches in personality assessment. In a special issue article, Lievens (2017a) outlined how SJTs might have great potential as alternative measures of personality and of the person-situation interplay. He argued that their use might extend from typical employment and educational settings to personality, clinical, and social psychology. Lievens also pointed out that SJTs might improve research on within-person variability, personality disorders, or personality facets. Thirty-two researchers in Personality, Social and Industrial/Organizational Psychology commented on this article and generally expressed their interest in using SJTs in these domains.

The objective of this chapter is to outline how SJTs might be adapted to measure personality traits, shed light on the person-situation interplay, and stimulate research on it. The structure of this chapter is as follows. First, we start with a brief definition and depiction of the past history of research on SJTs. Second, we describe the traditional view of SJTs as low-fidelity simulations. Third, we describe a more recent perspective that considers SJTs as alternative measures of personality. Fourth, we illustrate how SJTs might shed light on various forms of the personality-situation interplay.

**SJT: Definition and Brief History**

**Definition**

In Situational Judgment Tests (SJT), candidates are presented with short job-related situational descriptions and various response options to deal with the situations (Motowidlo et al., 1990). So, SJTs consist of two main components: a situation component and a response component. Regarding the situation component, candidates typically receive a brief description about the problem and some information about the people involved in the situation. Most SJTs still take the form of a written test because the scenarios are presented in a written format. Alternatively, in video-based or multimedia SJTs, a number of video
scenarios describing a person handling a critical job-related situation are developed (McHenry & Schmitt, 1994). At a critical “moment of truth”, the scenario freezes and applicants are asked to choose among several courses of action. Thus, video-based and multimedia SJTs allow the item context to be richly portrayed, thereby increasing their stimulus fidelity (Funke & Schuler, 1998; Olson-Buchanan & Drasgow, 2006). Recently, organizations are also exploring 2D-animated, 3D-animated, and even avatar-based SJTs (Weekley, Hawkes, Guenole, & Ployhart, 2015).

In the traditional SJT literature, situation descriptions (i.e., the item stems) are described as the heart of any SJT (e.g. Campion & Ployhart, 2013; Weekley et al., 2006). SJT theory postulates that situations in SJTs provide relevant context to applicants so that they can imagine themselves in a particular scenario and apply their context-dependent knowledge to respond to it. However, recent studies (e.g., Jackson, LoPilato, Hughes, Guenole, & Shalfrooshan, 2017; Krumm et al., 2015; Schäpers, Lievens, et al., 2019; Schäpers, Mussel, et al., 2019) have challenged this view and found that SJTs’ psychometric properties (e.g., construct saturation, SJT performance, or reliability) are not necessarily affected when situation descriptions were stripped off. For instance, Schäpers, Lievens et al. (2019) administered a video-based leadership SJT either with or without situation descriptions. The authors found that for between 47 and 88 % (uncorrected p-values vs. corrected for alpha inflation) of the video SJT items, it did not make a significant difference in SJT performance whether applicants worked on the version with or without scenarios. These findings underline that sufficient time and effort need to be invested in the development of the situation description. Otherwise, SJT situation descriptions might exert only limited impact (for further details see section SJT development). As a further explanation, one might argue that situation descriptions do not add much information and participants are able to construe the missing situation with the context given in response options (Melchers & Kleinmann, 2016). Thus,
more research is needed to investigate the role of situation descriptions in SJTs.

The response component exists in different formats. Upon reading the short scenarios, candidates might be asked to pick one response option from a list, rank the response options, or rate the effectiveness of these options. Contrary to self-report personality scales, candidates are not asked to rate themselves on the different response options. In SJTs, two instructions are typically distinguished: a knowledge-based instruction (“What should you do?”) and a behavioral tendency instruction (“What would you do?”). Knowledge-based instructions reflect maximal performance, whereas behavioral tendency instructions are related to typical performance (e.g., McDaniel et al., 2007). So, the instructions moderate the construct validity of SJT scores: SJT scores with knowledge-based instructions show higher correlations with general mental ability, whereas SJT scores with behavioral tendency instructions show higher correlations with Big Five personality (McDaniel et al., 2007). Notably, the moderating effect is particularly noticeable for Agreeableness, Conscientiousness, and Neuroticism. Furthermore, SJTs with behavioral tendency instructions are more prone to self-deception and impression management. However, the criterion-related validity of SJT scores is the same for both instruction formats (McDaniel et al., 2007).

Box 1 presents an example SJT item from the Teamwork-KSA Test (Stevens & Campion, 1999). The principle behind SJTs, similar to all simulations (e.g., assessment center exercises), is that applicants’ responses predict how they will behave in similar situations in the future. Commonly, work-related situations are developed from an analysis of critical incidents and are evaluated by subject-matter experts who determine what is an effective behavior for this situation at work (for an example see Weekley et al., 2006). For this item, answer B is the keyed answer because it reflects the most effective behavior, as determined by subject-matter experts. That is, these experts reasoned that disagreeable (but
routine) tasks should be fairly distributed within a team (regardless of task and position of the team members).

<table>
<thead>
<tr>
<th>Box 1. Example item from the Teamwork-KSA Test (see Stevens &amp; Campion, 1999)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suppose that you find yourself in an argument with several co-workers about who should do a very disagreeable, but routine task. Which of the following would likely be the most effective way to resolve this situation?</td>
</tr>
<tr>
<td>A. Have your supervisor decide, because this would avoid any personal bias.</td>
</tr>
<tr>
<td>B. Arrange for a rotating schedule so everyone shares the chore. <em>(correct answer)</em></td>
</tr>
<tr>
<td>C. Let the workers who show up earliest choose on a first-come, first-served basis.</td>
</tr>
<tr>
<td>D. Randomly assign a person to do the task and don't change it.</td>
</tr>
</tbody>
</table>

**Brief History**

SJTs are not new inventions. Early SJT versions go back to before World War II (WWII). During WWII, military selection psychologists needed an instrument 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 and they were asked which reaction they considered the most effective response (Northrop, 1989). Importantly, this instrument 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. In addition, the instrument enabled measuring recruits’ judgment skills in job-related settings, thereby significantly facilitating competent new soldier selection. This job test can be considered as one of the first 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 et al. reinvigorated interest in SJTs in the scientific community and in practice. Since then, SJTs
have become attractive selection instruments for practitioners who are looking for cost-effective instruments. Their use extends beyond employment settings because SJTs have also become popular as part of (medical) admissions and credentialing testing (e.g., Lievens & Sackett, 2012; Lievens, Sackett, & Buyse, 2009; Oswald et al., 2004; Patterson, Zibarras, & Ashworth, 2016; Prasad, Showler, Schmitt, Ryan, & Nye, 2017).

**The Traditional View: SJTs as Low-Fidelity Simulations**

**Underlying Theory and Rationale**

Traditionally, SJTs are seen as simulation-based instruments that are based on the behavioral consistency logic (Lievens & De Soete, 2012; Schmitt & Ostroff, 1986; Wernimont & Campbell, 1968). That is, the assumption that candidates’ performance during the selection procedure will be consistent with their future performance on the job. One of the main differences between SJTs and other simulation-based assessments relates to their level of fidelity. Fidelity can be conceptualized as the extent to which a selection task mirrors related job situations (Callinan and Robertson, 2000; De Soete & Lievens, 2015). Following this definition, some simulation-based measures (e.g., work samples or role plays) can be defined as high-fidelity simulations because applicants face actual problems (replicas of job tasks). On the other side, simulation-based measures such as SJTs are regarded as low-fidelity simulations (Motowidlo et al., 1990; Weekley et al., 2015) because they do not require test-takers to respond to job situations in reality, but instead confront them with written or video-based descriptions of realistic job situations. This also means that SJTs are more cost-effective than high-fidelity simulations and can be more easily administered to large groups of applicants. Hence, SJTs are often used as a pre-selection (“select-out”) tool, whereas high fidelity simulations are applied during later selection stages (“select-in”; De Soete & Lievens, 2015).
Using this logic, SJTs have been developed to capture domains as diverse as teamwork knowledge (McClough & Rogelberg, 2003; Morgeson, Reider, & Campion, 2005; Stevens & Campion, 1999), aviation pilot judgment (Hunter, 2003), team roles (Morgeson et al., 2005), emotion management (Blickle et al., 2009; MacCann & Roberts, 2008), call center performance (Konradt, Hertel, & Joder, 2003), proactivity (Chan & Schmitt, 2000), goal orientation (Westring et al., 2009), and academic performance (Oswald et al., 2004; Peeters & Lievens, 2005). Christian et al. (2010) developed a taxonomy to categorize these various domains. Their categorization showed that most SJTs capture leadership skills (37.5%), heterogeneous content (33.09%), and interpersonal skills (12.5%). In addition, SJTs were also developed to measure basic personality tendencies (9.56%), teamwork skills (4.4%), and job knowledge and skills (2.94%).

**SJT Development**

According to typical SJT guidelines, the development of SJTs includes three different stages (Corstjens, Lievens, & Krumm, 2017; Motowidlo et al., 1990; Ployhart & MacKenzie Jr., 2011). In the first step, test developers conduct a job analysis to generate different item stems (i.e., situation descriptions, see Flanagan, 1954). During this job-analysis, critical incidents of work situations are collected from subject matter experts (e.g., job incumbents or supervisors) that are representative of exceptionally good or exceptionally poor job-performance. Afterwards, the situations obtained are checked for level of specificity, redundancy, etc. The rationale behind this approach is to find and generate situations that are decisive for the success of the respective job.

In the second step, subject matter experts and/or (inexperienced) job incumbents generate different response options. An SJT item should contain both effective and ineffective behaviors for a respective situation. Subject matter experts are helpful for generating effective and ineffective reactions. Inexperienced employees might help to
generate a wider range of responses with different levels of effectiveness. For instance, inexperienced employees might be helpful for generating response options that may strike job occupants as reasonable, even if deemed less so by subject matter experts. Normally, an SJT-item contains four to five response options (Lievens, Peeters, & Schollaert, 2008). Furthermore, SJT developers have to decide on the response instructions (“would do” vs. “should do” instructions) and the response format. The latter can take a rate, rank, or most/least response format (for further information and effects on cognitive load and subgroup difference, see Arthur et al., 2014).

Finally, SJT developers decide on a scoring key. The following four categories are described in the SJT literature: (1) empirical, (2) theoretical, (3) rational/expert-based, and (4) hybrid scoring key. First, empirical scoring keys score response options according to their relationships with a criterion measure. Second, theoretical scoring keys follow a theoretical framework. This means that a reflected theory is used to determine (in-)effective behavior. Third, rational/expert-based scoring keys build upon subject matter experts’ ratings of the effectiveness of the different response options. Fourth, hybrid scoring keys combine at least two of the above-mentioned scoring keys. There is evidence that the scoring strategy might affect the psychometric properties of an SJT (e.g., magnitude of criterion-related validity; for further information see Bergman, Drasgow, Donovan, Henning, & Juraska, 2006), although different scoring keys also correlate often quite highly. For instance, there are indications that empirical, subject matter expert, and hybrid participation scoring keys might be the most effective keys for predicting work-related job performances, although more research is needed (see Bergman et al., 2006).

**Construct-Related Validity Evidence**

Similar to other simulations, SJTs present complex job situations that tap into multiple constructs. That is, to map the requirements of a job as well as possible, a collection
of different critical incidents is usually assessed. Although this approach may be conducive for predicting successful work performance, it challenges the internal psychometric analyses of SJT scores. Accordingly, SJTs show high item heterogeneity and low internal consistency (see, for example, Campion, Ployhart, & MacKenzie, 2014). The evidence of internal construct-related validity is often lacking (Guenole, Chernyshenko, & Weekly, 2017) because traditional SJT scores rarely break down in an assumed factor structure.

In the context of external construct-related validity, researchers extensively examined correlations between SJT scores and personality and cognitive ability (e.g., McDaniel et al., 2007). This is because personality and cognitive ability were often described as the main antecedents of SJT performance (Lievens & Motowidlo, 2016). While controlling for SJT content, meta-analytic results revealed that SJTs (especially those with a behavioral tendency response instruction) correlate moderately with scores on self-report personality scales (Agreeableness .37, Conscientiousness .34, Emotional Stability .35) and moderately with cognitive abilities (.35 with knowledge instructions; .19 with behavioral tendency response instructions). These meta-analytic results aggregate over many different types of traditional SJTs and provide only insight into how personality traits on average serve as antecedents on performing on the traditional SJT method in general. The moderate correlations can also be explained by the fact that traditional SJTs usually represent complex work situations and tap into many different domains. Therefore, a construct-driven SJT development has increased in popularity for some years now (see Guenole et al., 2017; Lievens, 2017b; for further details please see below).

**Criterion-Related Validity Evidence**

McDaniel et al. (2007) conducted a meta-analysis of the criterion-related validity of SJT scores in employment settings. They reported a mean corrected validity of .26. Furthermore, McDaniel et al. found that SJTs had incremental value for predicting job
performance over and above cognitive ability ($R^2$ increments .03 and .05; depending on response instructions) as well as self-rated personality ($R^2$ increments .06 and .07). So, although people’s SJT scores share variance with their self-report personality ratings, they are also sufficiently different that they have additional power to predict content-relevant behavior. Consistent with this, the same meta-analysis revealed that SJTs explained about 2% extra variance over and above both general personality and cognitive ability measures. Thus, SJTs increment can provide valuable information (especially for large organizations with large applicant pools). However, we would also like to stress the importance of not overinterpreting the results on incremental validity. In another meta-analysis, Christian et al. (2010) categorized the different domains that are captured in SJTs and compared their criterion-related validity regarding different job performance facets. The overall pattern of results showed relatively higher validities if the construct domain of the SJT matched the construct domain of the job performance facet. Finally, video-based SJTs appeared to show higher criterion-related validity than paper-pencil-based SJTs if interpersonal skills, leadership or heterogenous composites were captured. Due to this good criterion-related validity for predicting job performance (Christian et al., 2010; McDaniel et al., 2007) and favourable job applicant reactions (Kanning et al., 2006), SJTs have made strong inroads in selection in the last twenty years.

**SJT in the Training Context**

Although SJTs are predominately used as selection procedures, several researchers (e.g., Thornton III, Mueller-Hanson, & Rupp, 2017; Cox, Barron, Davis, & de la Garza, 2017) have argued that SJTs could also be useful for training purposes. As cost-effective simulations, SJTs might help trainees to learn how to handle critical work situations effectively. In contrast to other simulation-oriented approaches (e.g., unstructured critical incidents, role plays or case studies), SJTs offer various advantages due to their simple and
cost-sensitive application. For instance, SJTs allow trainees to put themselves in many different critical incidents and to receive feedback on performance immediately. Finally, SJTs also provide the opportunity to apply and test acquired knowledge. That is, SJTs might be useful to assess training effectiveness, especially if the evaluation of job performance is unavailable. As SJTs can be easily administered via an internet link, they serve as cost-effective pre- and post-measures (Ostroff, 1991).

Cox et al. (2017) provided initial evidence that SJTs are not only worthwhile in personnel selection but can also improve learning outcomes in a training context. They developed an SJT-based training module for managing a disaster sheltering center and found that trainees who participated in an SJT-based training module showed a greater performance improvement than trainees who participated only in a traditional lecture-based training. However, research on this topic is still in its infancy and comparatively few attempts have been made to explore the usefulness of SJTs for training purposes (for first examples see Cox et al., 2017; Hauenstein, Findlay, & McDonald, 2010). Therefore, additional research is required to fully understand to what extent SJTs are useful in a training context.

**SJT as Alternative Measures of Personality**

In the last few years, there has been a strong push to pay more attention to the constructs measured by SJTs and to adopt a more construct-driven approach for developing SJTs (Guenole et al., 2017; Lievens, 2017a, 2017b; Lievens & Motowidlo, 2016; Motowidlo, Hooper, & Jackson, 2006). This has given rise to the construction of SJTs that target (personality) traits (e.g., Bledow & Frese, 2009; Christian et al., 2010; Motowidlo et al., 2006; Mussel et al., 2018; Oostrom et al., 2019). One of the main differences from the classic SJTs is that the item stems (situations) of personality construct-driven SJTs are typically developed to activate a specific personality trait (e.g., Agreeableness) and the different response options represent different levels of the trait being targeted. For instance, some
options might reflect lower levels of Agreeableness, whereas other options indicate higher levels of Agreeableness. Recently, Guenole et al. (2017) outlined a stepwise procedure and an excellent set of recommendations for developing personality construct-driven SJTs that include the key features below. Table 1 presents the key differences between traditional SJTs and construct-driven SJTs (see also Lievens, 2017b).

Characteristics and Development

SJTs that measure personality traits have the following characteristics. First, the content (item stems and item options) of construct-driven SJTs are not necessarily developed with the help of subject matter experts. Although subject matter experts can still be of help, psychologists (e.g., graduate students familiar with personality trait psychology) are typically in charge of developing the item stems on the basis of trait activation theory (Tett & Burnett, 2003). This means that the situation depicted in the item stem is deemed relevant for eliciting the trait construct of interest. It is also possible to rely on situational taxonomies (see Gerpott, Thielmann, & Balliet, 2019; Parrigon, Woo, Tay, & Wang, 2017; Rauthmann et al., 2014) for developing the situations in the respective item stems. Generally, using either one of these approaches will make the item stem situations somewhat shorter, less contextualized, and more unidimensional (see Tett & Guterman, 2000 for examples of trait activation situations for various trait constructs).

As a second key feature, the item responses of construct-driven SJTs are more unidimensional because they operationalize different levels of the targeted trait instead of

---

1 There is conceptual overlap between construct-driven SJTs and the revealed trait technique (e.g., Costello, Wood, & Tov, 2018; Wood, Tov, & Costello, 2015). Similar to SJTs, the revealed trait technique confronts test-takers with scenarios and requests them to rate the likelihood of engaging in proposed actions. In addition, independent coders or subject matter experts rate the degree to that each action characterizes certain traits. Test-takers’ trait levels are then estimated by the correlation between the reported likelihood of engaging in certain actions and their representativeness for the respective trait. For more information, we refer to Costello et al. (2018) or Wood et al. (2015).
reflecting different behavioral categories. To generate the response options, SJT developers typically rely on groups of psychologists who propose and verify that the options reflect high or low levels of the traits, with results showing that this can be adequately done (see for example Motowidlo et al., 2006; Mussel et al., 2018).

Other differences are that construct-driven SJTs typically use a “What would you do?” response instruction and that a trait score is computed in the same way as in a personality scale by averaging (or summing) people’s endorsement of the response options (either high or low on the trait of interest). So, to score such SJTs, people receive higher scores when they endorse response options that are considered expressions of high levels of the specific trait targeted and vice versa. Apart from such trait scoring, another option is that people with considerable knowledge and experience with the particular contexts and subject matter experts involved score the effectiveness of the options (“effectiveness scoring”) and that candidates receive higher scores when their answers are similar to those of the experienced/ knowledgeable people.

Box 2 presents an example SJT item from Mussel et al. (2018) that aims to measure self-discipline. For this SJT, Mussel et al focused on a single facet within each Big Five dimension (instead of broad dimension such as leadership skills). As already noted above, the different item responses reflect different levels of the trait construct of interest instead of being indicative of a variety of constructs (as is the case in traditional SJTs).

<table>
<thead>
<tr>
<th>Box 2. Sample item from the facet self-discipline (see Mussel et al., 2018)</th>
</tr>
</thead>
<tbody>
<tr>
<td>You need to prepare a presentation at short-notice for an important meeting the next day. It's early evening and you're thinking about how to proceed. What would you do?</td>
</tr>
<tr>
<td>A. I stay longer in the office and finish the presentation. <em>(reflects higher levels of self-discipline)</em></td>
</tr>
<tr>
<td>B. After a short break I start working on the presentation so that I can finish it quickly. Afterwards, I enjoy the rest of the evening. <em>(reflects higher levels of self-discipline)</em></td>
</tr>
<tr>
<td>C. I relax in front of the TV after the stressful working day. Finally, I have enough time to finish the presentation later. <em>(reflects lower levels of self-discipline)</em></td>
</tr>
</tbody>
</table>
D. I am going to enjoy the evening and improvise in the meeting. (reflects lower levels of self-discipline)

Underlying Theory

Conceptually, construct-driven SJTs are posited to assess implicit trait policies (ITPs). ITPs are defined as inherent beliefs about the general effectiveness of actions that express traits to varying degrees (Lievens, 2017b, 2017a; Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo et al., 2006). So, ITPs reflect people’s procedural knowledge about relations between expressions of traits and their effectiveness in situations and people acquire them over time through general life experiences. Therefore, ITPs are viewed as characteristic adaptations instead of basic personality tendencies (see McCrae & Costa, 1996; DeYoung & Safron, this volume).

Implicit trait policies are not equivalent to personality traits because a person’s implicit trait policy represents his or her procedural knowledge about relations between expressions of traits and their effectiveness in situations to achieve some goal. The relationship between personality traits and implicit trait policies is grounded by the mechanism of dispositional fit (Motowidlo et al., 2006). That is, people’s personality traits interact with the traits expressed by the different response actions in the SJT in such a way that people who possess high levels of the trait expressed by the action believe that this action is more effective than people who have a lower standing on the trait. In addition, Motowidlo et al. (2006) argued that people who endorse SJT response options representing actions expressing specific traits will show corresponding trait-related behavior more often in real life. In other words, the notion of dispositional fit posits that by endorsing response options in SJTs people reveal something (1) about their standings on the personality traits represented by these response options, as well as (2) their likelihoods to show behavior expressing these associated traits in real life situations (Weekley et al., 2015).
Construct-Related and Criterion-Related Validity Evidence

Recent research with construct-driven SJTs provides evidence that it is possible to assess specific well-defined traits (i.e., people’s procedural knowledge about the effectiveness of these traits; e.g., Bledow & Frese, 2009; Motowidlo et al., 2006; Mussel et al., 2018). As one of the best examples, Mussel et al. (2018) developed an SJT for assessing five narrow traits (self-consciousness, openness to ideas, compliance, gregariousness, self-discipline). Their results indicated considerable promise for construct-driven SJTs. On average, facets assessed by the SJT obtained a convergent validity of .59 (from .41 to .70) with corresponding self-report ratings of these facets, while discriminant validity was -.01 (from -.31 to .19). The SJT facet ratings also predicted theoretically relevant outcomes over and above the self-report ratings of the same facets. Key “ingredients of their recipe” to successfully design such a construct-driven SJT included the reliance on clearly defined personality traits instead of vague dimensions and the development of response options to reflect low versus high trait levels. In addition, Mussel et al. developed a large set of items and measured each of the narrow traits with 22 items.

In several studies, Motowidlo et al. (2006) found support for their ideas regarding the role of ITPs. For example, they examined whether ITP scores for Agreeableness correlated with (1) self-reported Agreeableness ratings on the NEO-FFI and (2) behavioral expressions of Agreeableness as observed in actual situations. This was done via mini role-plays (e.g., interactions with coworkers, subordinates, supervisors, or customers). These role-plays were intended to activate Agreeableness but were not identical to the situations in the SJT. Independent observers rated the degree to which participants expressed trait-related behaviors in the role-plays. They found that ITP scores for Agreeableness significantly correlated with self-reported Agreeableness ($r$s between .29 and .31; depending on ITPs as correlational or
difference score) and significantly predicted Agreeableness scores in the role-plays (rs between .23 and .45) over and above self-reported Agreeableness.

Bledow and Frese (2009) developed an SJT for assessing personal initiative and investigated the construct-related and criterion-related validity of the SJT scores. They hypothesized that the assessment of personal initiative is compromised by Likert-type self-reports that ask for an abstract, generalized self-report because personal initiative is influenced by situational stimuli. Instead, Bledow and Frese (2009) argued that the situational character of personal initiative might be more appropriately measured via test-takers preferences for personal initiative across different situations. Therefore, they developed an SJT that confronts test-takers with different, concrete job situations and response options which mirror actions that differ in their level of personal initiative. In line with their hypotheses, their SJT showed a positive, but only moderate relation to Likert-type self-reports of personal initiative. Moreover, personal initiative as assessed by the SJT predicted a unique amount of variance in supervisor ratings of personal initiative and overall performance that was not explained by Likert-type self-reports. SJTs and Likert-type self-reports thus appear to measure different parts of the trait construct domain of personal initiative. An alternative explanation is that the SJT reduced the likelihood of socially desirable answers with the Likert-type self-ratings.

Furthermore, personal initiative as assessed by the SJT further mediated the effect of self-efficacy and felt responsibility on supervisor rated personal initiative and performance. Finally, personal initiative as assessed by the SJT converged with supervisor ratings of helping behavior, but not with supervisor ratings of conscientiousness. Thus, the pattern of convergent and discriminant relations indicates that the SJT measures test-takers’ preferences for personal initiative across different situations on the job.
As a final example, Oostrom et al. (2019) developed an SJT that aims to measure the HEXACO personality dimensions. Following the guidelines from Guenole et al. (2017) and Lievens (2017b), they developed four items for each of the six HEXACO personality dimensions. Overall, Oostrom et al. (2019) found positive evidence for SJT scores’ construct-related and criterion-related validity. For instance, SJT scores showed moderate to strong convergence with corresponding self- and other-reports. Further, SJT scores predicted criteria such as selection outcomes or supervisor rated organizational citizenship behavior. Honesty-Humility as assessed by the SJT was further found to predict supervisor rated organizational citizenship behavior after controlling for self-reports. In fact, Honesty-Humility as assessed by the SJT mediated the effect of self-reported Honesty-Humility on organizational citizenship behavior.

In sum, according to emerging evidence, construct-driven SJTs can be seen as a promising alternative to common personality assessments. In fact, they show satisfactory psychometric properties (i.e., satisfactory factor structure, reliability, and convergent validity) and explain incremental variance over and above self-reported personality.

**SJT as Alternative Measures of Personality and the Person-Situation Interplay**

In the previous sections, we discussed the traditional simulation-based view on SJTs and showed how more recent versions have started to measure trait constructs (personality traits). In this section, we go one step further and outline various strategies as to how SJTs might be further adapted to shed light on the personality-situation interplay.

**Assessment of Situation Construal**

As noted, SJTs consist of a situation component as well as a response component. In line with interactionism (Campion & Ployhart, 2013), it is assumed that candidates make sense of the short contextual SJT descriptions and that this situational construal guides their subsequent response choice. So, SJTs can be extended for use as a novel approach for
assessing situation construal, which is of pivotal importance in promoting SJT use in the personality domain due to its role in many theories (e.g., CAPS, Mischel & Shoda, 1995). This might allow better explaining and understanding people’s behavior, cognitions, and feelings.

One option to assess situation construal in SJTs was demonstrated by Rockstuhl, Ang, Ng, Lievens, and Van Dyne (2015). Their SJT did not present a list of predetermined (Multiple Choice) response options to individuals. Instead, they measured situation construal in an open-ended fashion by asking people to write down how they construed the situation. As another option, one might apply a multiple-choice approach in which one set of response options depicts different alternatives on how to perceive the situation, whereas another set of response options describes different alternatives on how to react to the situation. Test-takers might then select a given situation construal that matches best their personal situation construal (see Lievens, 2017a). In this way, SJTs might serve as a viable tool to identify discrepancies between individuals’ subjective construal and the consensual construal of a norm group (see also the section on Assessment of Personality Disorders below). As Rockstuhl (2017) noted, such insights might advance research on the subjective versus objective view of situations (see also Rauthmann, Sherman, & Funder, 2015).

Wood and colleagues (Wood, Lowman, & Harms; 2017; Wood & Spain, this volume) proposed to go one step further. They introduced the notion of field representations to more comprehensively capture the psychological situation. People’s functional field representations of the situation refer to their expectancies of how various features of the situation covary with each other and across possible ways of acting in this particular situation. People’s functional field representation are therefore supposed to subsequently influence their choice for or endorsement of different ways to respond to the situation. In SJTs, people’s functional field representations can be evaluated by asking people to not just
submit the action they would choose but also to evaluate how their chosen action will affect several aspects of the situation. This can be valuable for discerning the reasons that people understand as making actions more or less effective responses to specific situations (see Lowman, Wood, & Harms, 2020; Wood, Lowman, Harms, & Spain, 2019; Wood, Spain, & Harms, 2017).

Assessment of Within-Person Variability Across Situations

Almost all SJTs compute people’s mean score across the various items. However, SJTs and their inherent situation and response components lend themselves well to study meaningful variation in situation-trait contingencies, because SJTs acknowledge that people might adjust their response depending on the situation. SJTs thus enable computing two different sets of scores: an average score on each personality trait across the different SJT situations as well as a score reflecting variation in personality scores across these situations. Through the use of SJTs this means that researchers are able to assess both traditional cross-situational consistency (as reflected in mean trait level SJT scores) as well as within-person variability across situations (as reflected in standard deviation SJT scores associated with each trait).

Thus, SJTs that target specific traits also permit examining whether people vary their response indicative of a specific trait across situations (Dalal, Bhave, & Fiset, 2014). Recently, Lievens et al. (2018) found that people’s within-person variability on sociability and dutifulness in responding across written SJT situations explained incremental variance over and above mean SJT scores and self-reports of functional flexibility in explaining peer ratings of performance in team projects. Importantly, the within-person variability as measured by the SJT was also related to actual personality state variability measured two years later via a 10-day experience-sampling study.
Viewed in light of the broader personality literature (Baird, Le, & Lucas, 2006; Erickson, Newman, & Pincus, 2009), computing an index of within-person variability associated with scores per trait construct across situations via SJTs is useful because the standardized SJT format enables researchers to hold the situation constant. This is typically not the case when within-person variability is observed across situations/occasions via experience sampling methods (Fleeson & Gallagher, 2009; Furr, 2009).

In recent years, various analytical approaches have been used to go beyond analyses of SJT mean scores. For example, Lievens et al. (2018) measured within-person variability in responding across written SJT situations via IRT Tree Models (Böckenholt, 2012; De Boeck & Partchev, 2012; Lang, Lievens, De Fruyt, Zettler, & Tackett, 2019; Tutz, 1990). This IRT approach enables distinguishing between (a) test-takers’ mean-trait level/latent trait, (b) test-takers’ trait variability, as well as (c) items’ mean-level/latent difficulty, and (d) items’ propensity to elicit variability. To gain insight into the structure and response process underlying SJTs other studies applied other IRT models (e.g., Thiessen-Roe, 2013; Zu & Kyllonen, 2012) and generalizability theory (e.g., Jackson et al., 2017).

Assessment of Situation-Trait Contingencies

Besides providing an index of within-person variability, one might also go one step further and examine how people’s response option choices systematically vary according to specific situational dimensions or cues. This permits exploring whether there is meaningful group and individual variation in situation-trait contingencies (Huang & Ryan, 2011). Such “if-then” patterns are inherent parts of theoretical frameworks such as the Cognitive Affective Personality System (CAPS, Mischel & Shoda, 1995).

For both the assessment of within-person variability and situation-trait contingencies, it is important that the situations depicted in the SJT are carefully selected so that they mirror different types of psychological situations and trigger individuals to construe different
situations in different ways. One strategy consists of developing the SJT situations on the basis of situational taxonomies (e.g., Parrigon et al., 2017; Rauthmann et al., 2014; Ziegler, 2014). For example, recently, Rauthmann et al. (2014) developed the DIAMONDS situational framework. These situational dimensions in this framework might be used to design a set of SJT item stems. In turn, SJT response options might reflect different levels of personality traits. In that case, it is possible to examine to what extent people's (between-person) or one person's (within-person) responses differ across and within situations that tap into a specific situational dimension. In addition, one might scrutinize how the inclusion of particular omnibus contextual elements (see the situation cues in the framework of Rauthmann et al., 2014) affects situation perception and SJT response option choice (see also Brown, Jones, Serfass, & Sherman, 2016).

Such broad situational taxonomies (e.g., DIAMONDS, CAPTION) might predominantly capture situations in terms of their trait activation potential for specific traits. Therefore, as another strategy, the development of SJT situations might profit from more domain-specific situational taxonomies. For instance, to sample different leadership situations, test developers might rely on the taxonomy of different leadership situations described in the Multiple-Linkage Model (Yukl, 2010). To sample the interpersonal domain, test-developers might use the Interpersonal Circumplex Model (e.g., Carson, 1969; Kiesler, 1983). Clearly, future research related to broad and domain-specific situational taxonomies will benefit the development of SJTs to assess within-person variability and situation-trait contingencies.

**Assessment of Proactive Transactions**

Traditionally, SJTs are linear. That is, all applicants receive the same set of predetermined item situations and item options in the same sequence. So, the presentation of items is not dependent on people’s responses to previous items. This linearity might be seen
as contrived because it does not match how actual situations unfold. As the prototypical SJT puts constraints on individuals’ propensities to select situations, shape existing situations, and create novel situations, it means that one only gets insight into personality reactivity (Judge, Hofmans, & Wille, 2017).

To avoid this drawback and shed light on personality-driven situation experience (Rauthmann, Sherman, Nave, & Funder, 2015), one might expand on the traditional SJT paradigm. That is, one might design SJTs so that an applicant’s response to a situation determines the next situation that is presented. Accordingly, applicants are confronted with the consequences of their choices. This modality implies that all applicants do not respond to the same items. Such SJTs are called “branched”, “nested”, or “interactive” SJTs (Kanning et al., 2006; Olson-Buchanan et al., 1998; see also Weekley et al., 2015). Conceptually, such non-linear/branched/game-like SJTs also enable going beyond the reactive transactions and digging deeper into proactive/manipulative/evocative transactions because people have the opportunity to enter, avoid, or shape specific situations (Baumert & Blum, 2017; see also Weekley et al., 2015).

For example, an SJT assessing extraversion might confront a test-taker with the situation of a new co-worker introducing himself to all coworkers in a large open office space. A possible response option indicative of higher levels of extraversion involves actively starting a conversation with the new colleague. Test-takers who endorse such a response option then receive an item that has them actively enter this conversation with the new colleague. In this new item, the new colleague mentions that he did not yet figure out where to get coffee. A possible response option reflective of higher levels of extraversion involves suggesting to have a short coffee break together, whereas a possible response option indicating lower levels of extraversion only mentions to the new colleague where to get coffee. In contrast, test-takers who did not endorse the response option to actively start a
conversation with the new colleague would have avoided this situation. Instead, they might have endorsed a response option indicative of lower levels of extraversion (e.g., sending out an e-mail to the new colleague or not approaching the new colleague at all). These test-takers might then be confronted with a follow-up SJT item whose response options mirror different ways of answering or not answering an e-mail response from the new colleague. This example illustrates how interactive SJTs might serve to assess how test-takers actively enter, avoid, or shape situations. Accordingly, interactive SJTs also connect to the recent trend of gamification for assessment, selection, and development purposes (see, for example, Armstrong, Sanchez, & Bauer, 2017; Collmus & Landers, 2019; Georgiou, Gouras, & Nikolaou, 2019; Weekley et al., 2015; see also Miller, Jeong, & Christensen, this volume). Future research might investigate whether more interactive SJTs would ultimately benefit validity and utility for assessment, selection, and development purposes. Until today, this remains a promising but yet untested avenue for future research.

Assessment of Behavioral Responses

Traditional SJTs do not require people to show behavior. SJTs are low-fidelity simulations of what people say they will do in different situations (instead of what they actually do). It is possible, though, to require respondents to enact their responses to a video-based situation in front of a webcam (e.g., Cucina, Su, Busciglio, Harris Thomas, & Thompson Peyton, 2015; Lievens, De Corte, & Westerveld, 2015; Lievens, Sackett, Dahlke, Oostrom, & De Soete, 2019). These webcam vignettes are then coded and rated by experienced assessors. Accordingly, a hybrid is created between SJTs and assessment center (i.e., low fidelity assessment center or high fidelity SJT). Interestingly, one might also compare the potential discrepancy between people’s procedural knowledge (assessed via a written SJT) and their actually expressed behavior (assessed via assessment center exercises or via a webcam SJT; Breil, Geukes, & Back, 2017).
Assessment of Narratives and Goals

Apart from assessing trait-based aspects of personality, there exist several options to widen the SJT paradigm to include two other levels of personality: motivations/goals and narratives. For example, one might ask people to identify the story (narrative) that they associate with the scenarios. That is, people might be instructed to add short notes about past experiences that are similar to the scenarios depicted in the SJT (see Dunlop & Horton, 2017). Such elaboration might also be used to reduce faking good on SJTs (see Lievens & Peeters, 2008). In an extreme case, the SJT then becomes nothing more than a patterned behavior description interview, in which people are prompted to present a short narrative that provides information on their cognitions, emotions, and motivations for choosing specific behaviors in a specific situation and the goals they aimed to achieve. People’s selection of the situations then also tells something about their personality that goes beyond the conceptualization of personality as *personality traits* (for overviews of different conceptual levels of personality, see Dunlop, 2015; McAdams, 1995).

Similarly, one might prompt people to write down the motivations/goals they have in the scenario. Although it is possible to design SJTs’ response options so that they represent different strategies in goal achievement, these expansions of the SJT paradigm probably make most sense in an open-ended format (e.g., oral or written). For example, blank lines after the response options could request people to explain why they choose or endorse a specific response option and which specific motivation or goal this action might best serve. Such open-ended responses might provide insights into fundamental motivations such as needs for achievement, power, or affiliation (e.g., McClelland, 1987), regulatory focus or approach and avoidance motivation (e.g., Higgins, 1997; McClelland, 1987), and performance or learning goal orientation (e.g., Dweck, 1999). Further, open-ended responses
might also shed a light onto test-takers’ beliefs about how actions represented by the different response options would affect situations and be construed by other involved actors.

**Assessment of Personality Disorders**

So far, SJTs have mainly been used in employment settings. Yet, they might also be employed in clinical settings for assessing personality disorders. They then serve as an addition (or even alternative) to structured clinical interviews, self-report questionnaires, and ambulatory assessment approaches (Lievens, 2017; Wright & Simms, 2016). SJTs allow assessing how patients interpret and respond to a set of standardized scenarios and permit obtaining relevant diagnostic information. Examples are that they might indicate (1) endorsement of overtly problematic response options, (2) too much rigidity in reacting to situations activating the same traits, (3) extreme within-person variability (incoherence), and (4) situation-trait contingencies that mismatch those of “normal” people. One might also add think-aloud procedures to SJT items to examine how people construe the situations (see section on Assessment of Situation Construal).

**Epilogue**

So far, almost all SJT research has been conducted in employment and educational contexts. However, this chapter shows that SJTs also offer a lot of potential for the personality psychology field and for shedding light on the person-situation interplay. The use of SJTs should also not be limited to assessing procedural knowledge about the costs and benefits of trait-related behavior in work-related contexts. Procedural knowledge about the effectiveness of trait-related courses of action is as important in other life domains such as relationship longevity, popularity among peers, emotion management, school achievement or health management. Moreover, this chapter illustrates that adaptations of SJTs might serve to assess people’s situation construal, within-person variability across situations as well as situation-trait contingencies, proactive transactions, behavioral responses, narratives and
goals as well as personality disorders. These broad application areas might further increase the applicability of SJTs in the broader field of psychology.

References

IGNITE: Current trends in research and application. Symposium conducted at the 32nd Annual Conference of the Society for Industrial and Organizational Psychology. Orlando, FL, USA.


SJTs as Alternative Measures of Personality and the Person-Situation Interplay


Miller, L., Jeong, D., & Christensen, J. Game-based testing (under review). Interdependence approaches to the person and situation. In D. Wood, S. J. Read, P. D. Harms, & A. Slaughter (Eds.), *Emerging approaches to measuring and modeling the person and situation*. Amsterdam, NL: Elsevier.


SJT as Alternative Measures of Personality and the Person-Situation Interplay

and Organizational Psychology. Advance Online Publication. https://doi.org/10.1111/joop.12297


http://dx.doi.org/10.1111/j.2044-8317.1990.tb00925.x


https://doi.org/10.1016/j.hrmr.2009.03.007

https://doi.org/10.1002/per.2119


Mödling, Austria: Schuhfried GmbH.


### Differences in Development of Traditional versus Construct-Driven SJTs.

<table>
<thead>
<tr>
<th></th>
<th>Traditional SJTs</th>
<th>Construct-driven SJTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensionality</td>
<td>Multidimensionality (heterogeneity) at the item level</td>
<td>Unidimensionality (homogeneity) at the item level</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>Usually low to moderate internal consistency (for longer SJTs)</td>
<td>Satisfactory internal consistency</td>
</tr>
<tr>
<td>Development of situations</td>
<td>Critical incident technique with input from job holders and SMEs</td>
<td>Input from psychologists and psychological theories (e.g., trait activation theory and situational taxonomies)</td>
</tr>
<tr>
<td>Development of response options</td>
<td>Job incumbents generate response options that might tap various constructs. Effectiveness-ratings by SMEs and psychologists.</td>
<td>Input from psychologists to develop item responses reflect different levels of the trait construct of interest.</td>
</tr>
<tr>
<td>Response instructions</td>
<td>Knowledge (“Should do”) and behavioral (“Would do”) instructions</td>
<td>Knowledge (“Should do”) and behavioral (“Would do”) instructions</td>
</tr>
<tr>
<td>Scoring strategy</td>
<td>Effectiveness scoring</td>
<td>Effectiveness and trait scoring</td>
</tr>
</tbody>
</table>

Figure 1.

*Publications on SJTs per year and citations per year from 1994-2018.*

*Note.* This figure was created based upon a Web of Science search for “situational judgment test(s)” (31/12/2018). The left y-axis refers to the number of publications per year whereas the right y-axis refers to the number of citations per year.