8-2020

# The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions

Philipp SCHÄPERS
*Singapore Management University*

Patrick MUSSEL
*Freie Universitat Berlin*

Filip LIEVENS
*Singapore Management University*, filiplievens@smu.edu.sg

Cornelius J. KÖNIG
*Universitat des Saarlandes*

Jan-Philipp FREUDENSTEIN
*Freie Universitat Berlin*

*See next page for additional authors*

# The Role of Situations in Situational Judgment Tests: Effects on Construct Saturation, Predictive Validity, and Applicant Perceptions

**Article** *in* Journal of Applied Psychology · September 2019

DOI: 10.1037/apl0000457

CITATION
1

READS
278

**6 authors**, including:

Philipp Schäpers
Singapore Management University
**11** PUBLICATIONS **4** CITATIONS
SEE PROFILE

Patrick Mussel
Freie Universität Berlin
**69** PUBLICATIONS **588** CITATIONS
SEE PROFILE

Filip Lievens
Singapore Management University
**289** PUBLICATIONS **9,542** CITATIONS
SEE PROFILE

Cornelius J. König
Universität des Saarlandes
**140** PUBLICATIONS **2,387** CITATIONS
SEE PROFILE

# The Role of Situations in Situational Judgment Tests: Effects on Construct Saturation, Predictive Validity, and Applicant Perceptions

Philipp Schäpers, Patrick Mussel, Filip Lievens, Cornelius J. König, Jan-Philipp Freudenstein, & Stefan Krumm

Philipp Schäpers and Filip Lievens, Lee Kong Chian School of Business, Singapore Management University, Singapore; Patrick Mussel, Jan-Philipp Freudenstein, and Stefan Krumm, Institute of Psychology, Freie Universität Berlin, Germany; Cornelius J. König, Department of Psychology, Saarland University, Germany

Part of this research was funded by the German Research Foundation (KR 3457/2-1). Preliminary results of this research were presented at the 10th conference of the International Test Commission and the 18th conference of the European Association of Work and Organizational Psychology. We acknowledge the help of Luca Kröger, Luca Haensse, Raphael Cuadros, and Alexandra Göbel in collecting part of the data.

Correspondence concerning this article should be addressed to Philipp Schäpers, Singapore Management University, Lee Kong Chian School of Business, 50 Stamford Road, Singapore 178899, Singapore. Email: pschapers@smu.edu.sg

*This is the final author version (before journal's typesetting and copyediting) of the following article:* Schäpers, P., Mussel, P., Lievens, F., König, C. J., Freudenstein, J.-P., & Krumm, S. (2019). The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology.*

**Abstract**

Recent theorizing and empirical evidence suggesting that situational judgment tests (SJTs) are more context-*in*dependent than previously thought has sparked a debate about the role of situation descriptions in SJTs. To contribute to this debate and add to our understanding of how SJTs work, this paper conceptually embeds SJT performance in a situation construal model and examines the effects of situation descriptions on the construct saturation and predictive validity of SJT scores, as well as on applicant perceptions. Across two studies ($N$ = 1,092 and 578) and different SJTs, personality and cognitive ability were equally important determinants of SJT performance regardless of whether situation descriptions were presented or omitted. The effects of removing situation descriptions on the criterion-related validity of SJT scores differed depending on the breadth of the criteria. For predicting global job performance criteria (in-role performance and OCB), SJT validity was not significantly affected, whereas it decreased for predicting more specific criteria (interpersonal adaptability, efficacy for teamwork). Finally, the effects of omitting situation descriptions in SJTs on applicant perceptions were either negligible or small. Implications for SJT theory, research, and design are discussed.

*Keywords:* Situational Judgment Test, validity, contextualization, situation construal

# The Role of Situations in Situational Judgment Tests: Effects on Construct Saturation, Predictive Validity, and Applicant Perceptions

Many everyday work situations (e.g., a discussion with a supervisor, a customer complaint) require individuals to make an ad-hoc evaluation of the situational demands and then decide how to best respond. Simulating these processes in a low-fidelity format, situational judgment tests (SJTs) consist of job-related situation descriptions to which participants have to react by selecting, ranking, or rating multiple-choice response options (McDaniel, Hartman, Whetzel, & Grubb 2007; Motowidlo, Dunnette, & Carter, 1990). While the term "situational judgment" suggests that people's responses are more effective when they consider the specific demands of the situation, several recent studies (e.g., Krumm et al., 2015) have questioned the relevance of situation descriptions for SJT performance: For a substantial proportion of SJT items, performance was not affected when no situation descriptions were presented to test-takers.

This apparent discrepancy between long-held beliefs about SJT functioning and recent findings on the (ir)relevance of situation descriptions for SJT performance has sparked a vivid debate. Some scholars have called for directing more efforts to examining and developing generic and therefore cost-effective SJTs (e.g., Crook, 2016; Harvey, 2016), whereas others have been more prudent and raised several crucial albeit yet unanswered questions about the relevance of situation descriptions for SJTs' validity and their appeal to applicants.

This study adds to the current debate by conducting a more comprehensive investigation of the role of situation descriptions in SJTs. Such an investigation is pivotal because any conclusions about the relevance of situation descriptions in the SJT paradigm have to be drawn and balanced in light of the advantages (e.g., adequate validity, favorable applicant perceptions) that made SJTs popular in selection practice. Therefore, we aim to present evidence on the effects of situation descriptions on (a) the criterion-related validity of SJT scores, (b) the construct saturation of SJT scores, and (c) applicant perceptions.

## Study Background

### The Traditional View on SJTs

SJTs have traditionally been defined as low-fidelity simulations which "represent contextualized selection procedures that psychologically or physically mimic key aspects of the job" (Lievens & De Soete, 2012, p. 384). Similar to other simulations (e.g., assessment center exercises), SJTs build on the notions of point-to-point correspondence between simulated content (i.e., situation descriptions) and the criterion (future job situations) as well as on behavioral consistency (Bruk-Lee, Drew, & Hawkes, 2013; Lievens & De Soete, 2012). Table 1 (column 2) presents an example of a traditional SJT item. According to this traditional perspective on SJTs, by envisioning the simulated situation, participants should be able to make judgments about alternative ways of responding that mirror the judgments they would make in the real world (Campion & Ployhart, 2013). Thus, situation descriptions lie at the heart of SJTs because they simulate job situations and enable candidates to imagine themselves in a particular situation.

It has typically been taken for granted that individuals' judgments in SJTs depend on a thorough consideration of the situation. Only recently have direct tests of this assumption been conducted. Across a series of studies, Rockstuhl, Ang, Ng, Lievens, and Van Dyne (2015) asked participants not only what they would do in a (video-based) SJT situation, but also how they actually perceived a SJT situation (e.g., they had to make judgments about the actors' feelings and intentions). Their results revealed that appropriate situation construal was only a significant predictor of task and contextual performance when test-takers were asked to judge the situation. In addition, appropriate situation construal was substantially correlated with the respective SJT score. Other evidence in support of the traditional view on SJTs is provided by Westring et al. (2009). They decomposed response patterns on an SJT into trait-related and situation-related variance and found that situation factors accounted for 43% of the variance on average. By contrast, only 14% of SJT variance was explained by a trait-

related factor. In sum, these findings speak in favor of the notion that situations play an important role for SJT performance.

**Challenges to the Traditional View on SJTs**

Several recent studies (Jackson, LoPilato, Hughes, Guenole, & Shalfroosan, 2017; Krumm et al., 2015; Schäpers, Lievens, & Krumm, 2017), however, put a crack in the SJT edifice because they suggest that the role of situation descriptions has been overstated. Krumm et al. randomly assigned two versions of a team knowledge SJT (with and without situational descriptions in the item stem) to both students and employees (see columns 3 and 4 of Table 1, for an example SJT item without situation description). Remarkably, exclusion of situation descriptions in item stems did not make a significant difference in SJT scores for between 46% (when no correction was applied to the alpha level for making multiple comparisons) and 71% of the items (when the alpha level was corrected). These findings were replicated with different response instructions ("should do" vs. "would do") and SJTs from different construct domains (applied social skills, basic personality tendencies, and also job knowledge and skills). Furthermore, Schäpers et al. (2017) revealed that these findings even hold for situation descriptions presented in a video-based format. Finally, Jackson et al. (2017) found that situation-related effects explained only a small part of the reliable variance in SJTs. In fact, the largest proportion of variance could be attributed to a general performance factor. Thus, these recent studies suggest that SJT performance is more context-independent than context-dependent, which runs counter to the traditional SJT paradigm.

**The Debate**

Understandably, these results have sparked a vivid debate among researchers and practitioners (e.g., Lievens & Motowidlo, 2016; Melchers & Kleinmann, 2016; Naemi, Martin-Raugh, & Kell, 2016; Whetzel & Reeder, 2016). On the basis of this reduced importance of situation descriptions, some scholars stated that these results "lay the groundwork for developing cost-effective, off-the-shelf SJTs that can be used in a wide range

of occupations" (Harvey, 2016, p. 64) and thus more generic SJTs (e.g., Crook, 2016; Harvey, 2016). Conceptually, they also argued that SJTs should be better viewed as tests of general domain knowledge and that "a name change is in order" (Crook, 2016, p. 61).

Conversely, other researchers were more skeptical and added caveats to these implications, suggesting that it is premature to dismiss situation descriptions in SJTs (e.g., Chen, Fan, Zeng, & Hack, 2016; Fan, Stuhlman, Chen, & Weng, 2016; McDaniel, List, & Kepes, 2016; Melchers & Kleinmann, 2016). Specifically, it was argued that research on the effects on validity is needed because scores derived from SJTs without situation descriptions might exhibit lower validity, which would highlight that situation descriptions and situation perception are indeed integral parts of the SJT paradigm (e.g., Chen et al., 2016; Fan et al., 2016; Melchers & Kleinmann, 2016). Additionally, some scholars pointed to the relevance of situation descriptions for SJTs to engender favorable applicant perceptions (e.g., Crook, 2016; Fan et al., 2016). For example, Crook argued that "removing the job-specific situations may reduce favorable applicant reactions and the ease with which these measures will be embraced by managers for selection purposes" (p. 61).

A third group of reactions took a "middle-of-the-road" position (e.g., Harris, Siedor, Fan, Listyg, & Carter, 2016; Brown, Jones, Serfass, & Sherman, 2016; Ziegler & Horstmann, 2017). These scholars found the results not to be surprising because they suggest that individual differences such as personality and cognitive ability (and not the situations) are the main drivers of SJT performance. For example, Brown et al. (2016) posited that "personality is driving the behavioral response irrespective of the situation" (p. 41). Similarly, Ziegler and Horstmann (2017) argued that "this phenomenon underscores the overlap between SJTs and cognitive ability" (p. 46). Regardless of the perspective taken, there was general agreement that investigation into the role of situations and situational judgment in SJTs is needed to address several key unresolved issues.

## Unresolved Issues

**The role of situation descriptions for SJT validity and construct saturation.** As a first key issue, it is unclear whether performance differences between SJTs with vs. without situation descriptions translate into validity differences (see Chen et al., 2016; Fan et al., 2016; Melchers & Kleinmann, 2016). As argued above, it is also an open question whether the role of individual difference variables (such as personality, cognitive ability, etc.) become more or less important determinants of SJT performance when respondents can no longer use the item stem information to form their situation perception (Harris et al., 2016; Sherman, Rauthmann, Brown, Serfass, & Jones, 2015). In other words, how does the cognitive and personality saturation of SJT scores change for SJTs with or without situational descriptions?

To answer these questions, this study draws, this study draws upon situation construal models (e.g., Block & Block, 1981; Funder, 2016; Hogan, 2009; Mischel, 1977; Rauthmann et al., 2015; Reis, 2008). Situation construal is defined as a person's distinctive perception of the situation (i.e., the psychological situation, Block & Block, 1981; Funder, 2016; Mischel & Shoda, 1995; Rauthmann et al., 2014) that is determined by person variables as well as the objective situation (i.e., the situation as agreed upon by many people; Block & Block, 1981).

***Situation construal model of SJTs with situational item stems***. Although everybody typically receives the same SJT situations, situation descriptions in SJT items are ambiguous because they are short and do not present all of the contextual information. People might therefore interpret them in distinct and unique ways (Harris et al., 2016; Melchers & Kleinmann, 2016; Meyer, Dalal, & Hermida, 2010; Mischel, 1973; Sherman et al., 2015). This is why situation construal is relevant in SJTs, even though it is not explicitly measured (SJTs ask people only what they would/should do instead of how they perceive the situation; Rockstuhl et al., 2015). We posit that people's differential perceptions of SJT item situations result from the interaction of people's personality and the objective situation (see also Sherman et al. 2015). Depending on the SJT situation given and one's personality, we propose

that each individual construes the SJT situation in a unique way (Allport, 1961; Reis, 2008; Sherman et al., 2013).

Figure 1a presents our adaptation of the situation construal model to SJTs. The objective situation side refers to the situation in the item stem that is presented to all test-takers. On the person side, we include personality and cognitive ability, which represent important antecedents of people's procedural knowledge (Motowidlo & Beier, 2010). We posit that these individual difference variables might influence how people construe the situation. Recently, Sherman et al. (2013) confirmed this link between personality and situation construal of everyday situations for all Big Five factors. For example, people high on Agreeableness tended to construe situations more as opportunities to get along with others and cooperate, whereas people high on Openness tended to perceive situations as more intellectually stimulating (see also Serfass & Sherman, 2013; Sherman et al., 2013).

The last part of the model focuses on people's responses to the SJT situation. It is posited that situation construal drives people's SJT responses (together with main effects of the person and the situation). Hence, in this model situation construal is a precursor of successful SJT responding (Harris et al., 2016). Finally, people's responses to SJTs—along with their personality, cognitive ability, procedural knowledge, and other personal qualities (Barrick, Mount, & Judge, 2001; Schmidt & Hunter, 1998)—should be predictive of subsequent real-world behavior and performance (captured via criterion measures; Christian, Edwards, & Bradley, 2010; McDaniel et al., 2007).

***Situation construal model of SJTs without situational item stems.*** Viewed in light of the situation construal model, removing situation descriptions from SJTs means that the objective situation is taken out. As shown in Figure 1b, when one determinant of situation construal, the objective situation, is no longer present, the effects of the other determinant, the person variables, can be expected to increase. So, the absence of a situation description will

alter the construct saturation[1] of SJT scores. Specifically, it will make SJT performance more saturated with person-based antecedents. In this study, we focus on personality and cognitive ability (see Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010).

More specifically, we expect the effect of personality traits on SJT performance to increase for two reasons. First, in Figure 1a, personality has both a direct and an indirect effect (through situation construal) on SJT response choice. As we can assume the indirect effect to be heavily diminished in the absence of a situational stem, the direct effect on response choice will increase in importance. Or as Harris et al. (2016) put it: "In the absence of situational cues, [...] SJT performance is increasingly a function of personality traits (reflecting the idea of personality as a generalized representation of how persons behave across situations)" (p. 25; see also Sherman et al., 2015).

Second, the increasing role of personality when no situation descriptions and only response options are presented is in line with the notion of dispositional fit (Motowidlo, Hooper, & Jackson, 2006). That is, people's traits interact with traits expressed by the different SJT response options in such a way that people who possess high levels of the trait expressed by the response action believe that this action is more effective than people with lower levels of the trait. So, the correlations between people's SJT scores and their ratings on corresponding personality traits will be larger when situation descriptions are absent. Thus:

*H1: The personality saturation of SJT scores will be significantly higher for SJTs without situation descriptions than for SJTs with situation descriptions.*

In addition to increased personality saturation, we also anticipate the role of cognitive ability in determining SJT performance to increase in the absence of situation descriptions. For typical SJTs with situation descriptions, the meta-analytic correlation between cognitive ability scores and SJT performance is moderate ($\rho = .32$; McDaniel et al., 2007). However,

[1] Construct saturation refers to the degree to which total score variance in a measure reflects specific construct variance (Dahlke & Sackett, 2017; Lievens & Sackett, 2017; Lubinski & Dawis, 1992; Roth, Bobko, McFarland, & Buster, 2008).

the confidence interval is large (.08-.57), indicating that there is a lot of variability in cognitive saturation across SJTs. We posit that the inclusion/exclusion of a situational item stem might function as a yet unexamined moderator. Although people are required to read less text in the absence of a situation description in the item stem, a key point is that the SJT item becomes more difficult to solve because people lack pieces of information to guide their response choice (Ziegler & Horstmann, 2017). We posit that when facing SJT items without situational item stems, people high on cognitive ability will be better able to "fill in the holes" and deduce correct responses solely on the basis of the response options (cf. Vickers, Mayo, Heitmann, Lee, & Hughes, 2004; Vernon, & Strudensky, 1988). Therefore:

*H2: The cognitive ability saturation of SJT scores will be significantly higher for SJTs without situation descriptions than for SJTs with situation descriptions.*

In sum, Figure 1a and 1b show that response choice in SJTs with and without situations has different determinants. Whereas performance in SJTs with situation descriptions captures the direct effects of person variables and of situation construal, performance on SJT items without situation descriptions primarily reflects the direct effects of individual differences (see also Harris et al., 2016; Lievens & Motowidlo, 2016; Sherman et al., 2015; Ziegler & Horstmann, 2017). Given that construct saturation can mediate the effects of predictor method factors (e.g., the presence of situation/context descriptions) on validity (Lievens & Sackett, 2017), the next question is how these different drivers of SJT performance with and without situation descriptions translate into validity differences.

A feature of Figure 1a and 1b is that they are nested. That is, one (situation description in the item stem) of the two determinants of response choice presented in Figure 1a is removed in Figure 1b. Therefore, comparing the validity of SJT scores across the two conditions is a test of whether situation construal has added value to increase the predictive power of SJT scores. Thus, as explained below, the question of whether the validity of SJT scores is affected when their item stems no longer contain situation descriptions comes down

to a test of the assumptions underlying interactionism (Rauthmann et al., 2015). That is, if the validity of SJT scores without situation descriptions is significantly lower than the validity of SJT scores with situation descriptions, this suggests that situation construal has incremental predictive power in how it interacts with traits to determine response choice (see Figure 1a). Such a result conforms to interactionist models (Campion & Ployhart, 2013; Mischel & Shoda, 1995; Tett & Burnett, 2003) that posit best predictions are obtained on the basis of how people's traits interact with the situation. Several scholars have indeed argued that there will be a reduction in validity when the role of situation construal is muted due to the absence of situational item stems (Chen et al., 2016; Fan et al., 2016; Melchers & Kleinmann, 2016). A finding of validity decrease for SJT scores without situation descriptions lends also support to the notion that SJT scores tap more into context-dependent knowledge.

Conversely, if there is no significant difference between the criterion-related validity of SJT scores with and without situation descriptions, this suggests that situation construal has little added value to increase prediction. In that case, SJT performance that is determined primarily by traits, abilities, and knowledge suffices to predict future performance. Such a finding suggests that the best predictions are obtained from people's generalized tendencies to react and is thus not consistent with interactionism. So, a finding of no criterion-related validity difference between SJT scores with and without situation descriptions also supports that SJT scores tap more into context-independent knowledge. As the answer to the question of the effect of situational item stems in SJTs on criterion-related validity depends on the conceptual perspective (interactionist or not), we put forward a research question:

*Research Question 1 (RQ1): Is the criterion-related validity of SJT scores affected when respondents no longer receive situation descriptions in SJT item stems?*

**The relevance of situation descriptions for applicant perceptions.** Generally, studies have shown that SJTs lead to favorable applicant perceptions (e.g., Chan & Schmitt, 1997; Salgado, Viswesvaran, & Ones, 2001; Whetzel & McDaniel, 2009). Apart from effects

on criterion-related validity, various scholars noted that reducing or even stripping the item stems from SJTs might lower the appeal of SJTs among applicants and users (e.g., Crook, 2016). One reason is that the face validity and perceived predictive validity of SJTs is reduced because the situations in the item stems typically reflect actual job situations. It might thus be less obvious how a stemless SJT relates to judging situations and responding to them in a particular job (Crook, 2016). Second, stemless SJT items are even more ambiguous and vague than prototypical SJT items, which makes it difficult for applicants to engage in situation construal (McDaniel et al., 2016). The items might also be perceived as incomplete. Accordingly, it becomes more challenging for applicants to gauge whether they scored well on them. This would then lower ratings on fairness dimensions such as opportunity to perform and perceived knowledge of results. Third, applicants might also become increasingly frustrated by the difficulty to make sense of the items due to the lack of information inherent in stemless SJT items, leading to negative affect and reduced test-taking motivation. Given that these aspects are known to negatively influence applicant perceptions (Schmitt & Gilliland, 1992; Hausknecht, Day, & Thomas, 2004), we posit:

> *H3: Applicants' perceptions of procedural fairness dimensions (face validity, perceived predictive validity, opportunity to perform, and perceived knowledge of results), positive affect (enjoyment), and test-taking motivation will be higher for SJTs with situation descriptions in the item stem than for SJTs without situation descriptions in the item stem.*

## Present Studies

To test our hypotheses and research question, we conducted two studies that manipulated the presence or absence of situation descriptions in the item stems of three SJTs. In particular, Study 1 tested differences in personality saturation (H1), cognitive ability saturation (H2), and applicant perceptions (H3), whereas Study 2 addressed H1 and H2 as well as the Research Question about effects on validity. In Study 1, we used the SJT on

personal initiative (Bledow & Frese, 2009) and the Situational Judgment Test for Teamwork (Gatzka & Volmer, 2017). Study 2 used the Team Role Test (Mumford, Van Iddekinge, Morgeson, & Campion, 2008).

## Study 1

### Methods

**Participants and procedure.** An a-priori power analysis with G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) revealed that a sample size of $N = 572$ participants is necessary to detect even small differences in correlation coefficients between two groups (e.g., with vs. without situation descriptions) with sufficient statistical power ($1 - \beta = .80$). Participants were recruited via a scholarly-hosted online panel that consisted of individuals who had declared their willingness to participate in psychological research. This panel includes both students (20%) and working people (52%); covering a wide variety of educational levels (e.g., 25% university degree, 30% A-Levels, 29% O-Levels). The average age is 47.08 years ($SD = 14.45$; for further information, see Göritz, 2014; Göritz, Borchert, & Hirth, 2019). As compensation for their participation, participants received feedback on their performance. Approximately 14,000 subscribers were contacted via e-mail (response rate = 12.9%). To minimize participant burden and drop-out, we administered this study's tests across three different sessions. The time between the test sessions varied between 2 and 3 weeks. Content and design of Study 1 are similar to Study 2, which was approved by the Institutional Review Board of the Freie Universität Berlin (No. 88/2014) as part of a funded research project granted to the last author (German research foundation DFG; KR 3457/2-1). Following recommendations by Meade and Craig (2012), we checked for irregular responding (bogus items, instructed response items, and self-declaration of data exclusion below) and excluded 145 (11.7% of the initial sample) participants from further analyses.

The final sample of $N = 1{,}092$ participants (56.4% female) with a mean age of 51.05 years ($SD = 13.72$, range from 18 to 87) completed at least two of the three test sessions. The

majority (67.2%) of the final sample were working people covering a wide variety of different educational levels. Specifically, 32.5% held a university master's degree, 22.4% held a university entry qualification (A-level), and 30.0% held a tenth-grade degree.

**Study design and materials.** All data were collected online and followed a three-step procedure. First, participants completed a cognitive ability test and responded to several demographic questions. In session two, participants completed the SJT on personal initiative. After the SJT, applicant perceptions were assessed via six different scales (see below). In session three, the teamwork SJT was administered. Again, applicant perceptions were assessed upon SJT completion. Following recommendations by Osborne and Overbay (2004), individuals were excluded from further analyses if *z*-scores were below -3 or above +3, which has been shown to lead to more accurate estimates.

We relied on a between-subjects design to test our hypotheses. Participants were randomly assigned to one of three conditions. In the first condition, the aforementioned SJTs were administered with situation descriptions. In the second condition, participants received the same SJTs, but without situation descriptions (i.e., item stems were omitted). Although these two conditions constituted our main conditions, we also followed up on the suggestion of an anonymous reviewer by adding a third condition to our study design. In this condition, we not only omitted situation descriptions but also eliminated references to context (which referred to the previous item stem) from the response options. This also avoided awkward phrasing in response options, which might suggest to test-takers that some part of the item had been deliberately omitted (for an example, see column 4 of Table 1). Hence, in a third condition, we changed the wording of the response options if needed to make them appear less incomplete (artificial) and more generalized so that they still made sense.

***Situational judgment tests (SJTs).*** The SJT on personal initiative (Bledow & Frese, 2009) and the teamwork SJT (Gatzka & Volmer, 2017) both consist of 12 descriptions that refer to critical incidents concerning personal initiative or teamwork. Both SJTs were

administered with behavioral tendency response instructions ("what would you do?") and a pick-the-best response format with four to five different response options. Test-takers gained one point per correct response choice and lost one point if they selected an ineffective response option. The sum of all points was used as SJT score.

***General mental ability.*** We used the short version of the Hagen Matrices Test (six items, Heydasch, Haubrich, & Renner, 2013) to measure participant's cognitive abilities. Each item consists of a 3 × 3 matrix with one missing field. Participants were asked to complete the missing field correctly. Participants were given a time limit of 2 minutes per item. The number of correctly answered items was used as the score (1 or 0 points for each item). The reliability of the test was satisfactory; $\omega^2$ ranged from .71 to .76 (see Table 2).

***Big Five personality dimensions.*** The Big Five personality dimensions were assessed with a short version of the Big Five Inventory (BFI-K; Rammstedt & John, 2005). Participants responded to 21 items on a 5-point rating scale (from 1 = *disagree strongly* to 5 = *agree strongly*). Reliability (ω) of the BFI-K ratings ranged from .59 to .86 (see Table 2). The online panel database provided self-ratings on this Big Five personality questionnaire for 300 participants in the final sample. This means that these data were collected in the study of Heidemeier and Göritz (2016) via a similar design as ours (i.e., use of an online questionnaire with several measurement points, with test-takers participating on a voluntary basis).

***Applicant perceptions.*** We assessed applicant perceptions with six different measures. More specifically, we used four measures from Smither, Reilly, Millsap, Pearlman, and Stoffey (1993) that were adapted to reflect the specific SJTs (teamwork or rather personal initiative): Face validity (five items; e.g., "I did not understand what the examination had to do with working on a team"), perceived predictive validity (five items; e.g., "Failing to pass the examination clearly indicates that you can't work on a team"), perceived knowledge of

[2] Especially in SJTs, the assumption of tau-equivalent models (i.e., equal factor loadings) is typically not met. Thus, we calculate McDonald's Omega as it is a more appropriate estimate for reliability in such cases (compared to Cronbach's Alpha; see Dunn, Baguely, & Brunsden, 2014).

results (three items; e.g. “After I finished the examination it was clear to me how well I performed”), and positive affect (two items; “I enjoyed the examination to a great degree”). We also measured applicants’ perception of their chance to perform (4 items, Bauer et al., 2001; e.g., “I could really show my skills and abilities through this test”) and test-taking motivation (five items, Arvey, Strickland, Drauden, & Martin, 1990; e.g., “I wanted to perform well on the tests”). Ratings were made on a 5-point rating scale (1 = *disagree strongly;* 5 = *agree strongly*). Reliability (ω) of the scales’ ratings ranged from .61 to .88 (see Table 2).

***Careless responding.*** Following recommendations by Meade and Craig (2012) to detect careless responding, we added two bogus items to session one (Anderson, Warner, & Spencer, 1984; Carroll, Jones, & Sulsky, 2004; Levashina, Morgeson, & Campion, 2009): participants had to indicate their level of familiarity with non-existent subjects (e.g., “To what extent have you used Johnson’s dyadic approach of avoiding conflict in work teams?”; see Levashina et al., 2009). Participants were excluded from our analyses if they indicated they had used these non-existent techniques. In session two and three, we inserted two instructed response items (e.g., “To monitor quality, please respond with a two for this item” see Meade & Craig, 2012). Participants were excluded if they failed to answer these items correctly. Finally, we also asked participants whether their data could be used (Meade & Craig, 2012).

**Results and Discussion**

**Preliminary analyses.** We ran preliminary analyses to rule out alternative explanations for differences across the experimental conditions. Importantly, participants did not differ in terms of gender, $\chi^2(4) = 1.44$, $p = .84$, $\phi = .04$, age, $F(2, 1089) = 0.671$, $p = .51$, partial $\eta^2 = .001$, level of education, $\chi^2(10) = 17.70$, $p = .06$, $\phi = .13$, country of origin, $\chi^2(2) = .52$, $p = .77$, $\phi = .02$, or cognitive ability, $F(2, 1089) = 1..159$, $p = .31$, partial $\eta^2 = .002$. Next, we tested whether the reliability of the administered SJT scores differed between the conditions. Reliability estimates were consistent with meta-analytic estimates on SJT score

reliability (Campion, Ployhart, & MacKenzie, 2014; Catano, Brochu, & Lamerson, 2012; Kasten & Freund, 2016). That is, ω ranged from .68 to .73 (for the personal initiative SJT) and from .34 to .50 (for the teamwork SJT). Importantly, the SJT reliability estimates did not differ significantly between the conditions (see Table 2).

Finally, we performed multiple group measurement invariance analyses. Specifically, we tested for metric invariance (i.e., assuming equal factor loadings across groups), which is necessary to allow interpretations of between-group differences (Bollen, 1989). We used R Studio (version 1.0.143; R Core Team, 2016) and the R package lavaan (version 0.5–22; Rosseel, 2012). On the basis of several researchers' recommendations (Beauducel & Wittmann, 2005; Browne & Cudeck, 1993; Byrne, 1989; Hu & Bentler, 1999; Kline, 2004), model fit was considered acceptable when comparative fit index (CFI) was > .90, root-mean-square error of approximation (RMSEA) < .10 (preferably < .05), and standardized-root-mean square residual (SRMR) < .10. For the SJT on personal initiative, we randomly created four parcels of three items each (Little, Cunningham, Shahar, & Widaman, 2002). The fit did not differ significantly from the baseline model when restraining the factor loadings between the different SJT versions: $\chi^2(12) = 8.046$ ($\Delta\chi^2(6) = 4.81$, $p = .57$), RMSEA = .000 (90% CI [.000, .036]), SRMR = .021, CFI = 1.000. For the teamwork SJT, we specified the proposed factors of the test authors and created three parcels, wherein the composition of the parcels items was determined by the factor affiliation. Multiple group measurement invariance analyses revealed that model fit did not decrease substantially in comparison to the baseline model[3] when assuming equal factor loadings across the three groups: $\chi^2(4) = 1338$ ($\Delta\chi^2(4) = 1.338$, $p = .85$), RMSEA = .000 (90% CI [.000, .049]), SRMR = .011, CFI = 1.000, thus supporting metric invariance.

[3] Fit indices are not reported for this just identified model because "estimated parameters perfectly reproduce the sample covariance matrix, chi-square and degrees of freedom are equal to zero, and the analysis is uninteresting because hypotheses about adequacy of the model cannot be tested." (Ullman & Bentler, 2012, p. 665).

**Test of SJT score differences.** Similar to Krumm et al. (2015), we compared mean scores for SJTs with vs. without situation descriptions in the item stems. For the personal initiative SJT, we found a significant albeit small effect for the overall score: Test-takers receiving items with situation descriptions had higher scores than participants receiving items without situation descriptions, $t(714.981) = 2.36$, $p < .05$, $d = .18$. We followed up on these results with results on the item level. Four out of 12 items (33%) had significantly higher scores when situation descriptions were presented than when situation descriptions were omitted. When a correction for alpha inflation was applied (Cabin & Mitchell, 2000), significantly higher scores were obtained for 2 out of 12 items (i.e., 17%; see Table 3). For the teamwork SJT, test-takers who completed the SJT with situation descriptions also obtained a significantly higher SJT score than those who did not receive situation descriptions. This effect was large, $t(549) = 6.32$, $p < .01$, $d = .54$. At the item level, it made a significant difference whether the situation description was presented for between 50 and 67% of the items (depending on the correction for alpha inflation, see Table 3).

Very similar results were obtained when SJT items with situation descriptions were compared with SJT items with situation-neutral response options (i.e., with situational information converted into more general information). For the personal initiative SJT, 6 out of 12 items (also 6 out of 12 with the alpha inflation correction) had significantly higher scores when situational descriptions were included. For the teamwork SJT, 7 out of 12 items (4 out of 12 when correcting for alpha inflation) had significantly higher scores when situational descriptions were included.

**Test of construct saturation hypotheses.** To test H1 and H2, we examined differences in cognitive ability and personality saturation between the three SJT versions. We compared zero-order correlations of the three SJT versions with the cognitive ability test and self-rated personality (see Table 4). We applied Fisher's *z* transformation for all comparisons (Cohen, Cohen, West, & Aiken, 2003). No significant differences in the hypothesized

direction occurred for either the personal initiative or teamwork SJT, lending no support to Hypotheses 1 and 2. Similar results were obtained when SJT items with situation-neutral response options were compared to SJT items with situation descriptions (Table 4). The only exception formed the personal initiative SJT version with situation descriptions which showed a higher correlation with Neuroticism than the version with situation-neutral response options.

To further scrutinize the construct saturation between the three SJT versions and personality/cognitive ability, we specified a multi-group path model with general mental ability and personality predicting SJT performance (see Figure 2). To ensure model identification and parsimony, personality and cognitive ability were specified as manifest variables. The baseline model showed a good fit, $\chi^2(129) = 158.97$, RMSEA = .025 (90% CI [.007, .038]), SRMR = .051, CFI = .97. When restraining all relevant path coefficients to differ across the three groups, the model fit did not decrease significantly, $\chi^2(153) = 184.00$ ($\Delta\chi^2(24) = 25.725$, $p$ =.37), RMSEA = .024 (90% CI [.005, .035]), SRMR = .063, CFI = .97.

In short, personality saturation did not seem to be contingent on the availability of situation descriptions. Likewise, all SJT versions were equally correlated with general mental ability. So, situation descriptions neither add complexity (through the requirement to read and understand text) nor reduce complexity (by making judgments less ambiguous). Alternatively, one might argue that all SJT versions are similar in terms of cognitive load.

**Test of differences in applicant perceptions.** H3 stated that the absence of situation descriptions will negatively affect applicant perceptions. Table 2 shows that in all cases, applicant perceptions were descriptively higher in the condition with situation descriptions. To test our hypothesis across all six applicant perception measures, we conducted a one-way MANOVA per SJT. For the personal initiative SJT, results revealed a significant MANOVA, $F(12, 2168) = 1.885$, $p$ =.03; Wilk's $\Lambda = 0.979$, partial $\eta^2 = .010$. Post-hoc tests (Gabriel) indicated that face validity was perceived more positively in the condition with situation descriptions than in the condition with situation-neutral response options ($p < 05$). For the

teamwork SJT, we found a significant main effect for the availability of situation descriptions, $F(12, 1662) = 3.299$, $p < .01$, Wilk's $\Lambda = 0.954$, partial $\eta^2 = .023$. Post-hoc tests (Gabriel) showed for two of the six applicant perception dimensions (face validity and affect) more favorable applicant perceptions in the condition with situation descriptions than in the condition without situation descriptions. No significant differences were found for perceived predictive validity, perceived knowledge of results, opportunity to perform, or test-taking motivation. Similar results were obtained when SJT items with situation descriptions were compared with SJT items with situation-neutral response options. Post hoc tests (Gabriel) revealed that for face validity and opportunity to perform perceptions were more favorable in the condition with situation descriptions than in the condition with situation-neutral responses. Thus, as we found support for our hypotheses for only a couple of applicant perception dimensions, there was only partial support for H3. Overall, effect sizes were also small for the personal initiative SJT (from $d = |0.01|$ to $|0.17|$, mean = $|0.07|$) and small to medium for the teamwork SJT (from $d = |0.08|$ to $|0.41|$, mean = $|0.19|$).

In sum, presenting situation descriptions in SJTs had a positive but mostly small effect on two applicant perceptions dimensions (face validity and affect) for the teamwork SJT. However, no such effect was found for the remaining four applicant perception dimensions. For the personal initiative SJT, applicant perceptions did not differ between SJTs with or without situation descriptions. Notably, almost the same results were obtained when SJT items without situation descriptions and situation-neutral response options were compared to SJT items with situation descriptions, with the exception of opportunity to perform. This suggests that the potential side effect of awkwardly-phrased responses due to references to a missing context did not greatly affect applicants' perceptions. Thus, contrary to recent statements (e.g., Crook, 2016), situation descriptions may be at best only slightly relevant for ensuring favorable applicant perceptions.

Study 1 examined the relevance of situation descriptions for SJT scores' construct saturation (personality and cognitive ability) as well as for applicants' perceptions to SJTs. We conducted a second study using a third SJT, the Team Role Test (Mumford et al., 2008), to identify if the effects of manipulating situation descriptions on test score differences and construct saturation replicate and generalize. We also scrutinized the effects on criterion-related validity. Study 2 did not include the situation-neutral condition since this condition yielded results that were virtually identical to the without situation condition.

## Study 2

### Methods

**Participants and procedure.** We recruited participants via online postings (on Facebook, university websites, and in newsletters), poster advertising, or actively approached them on the campus of a large German state university (see Study 1 with a similar design for an a-priori determination of sample size to have sufficient statistical power). Inclusion criteria for participation were experience in teamwork, fluency in German, and an age of 18 years or older. After excluding 26 participants (due to insufficient German language skills or failure to respond honestly to the bogus items, see below), the actual sample consisted of 578 participants (68.2% female). The sample included both students (59.7%) and non-students (40.3%), with the latter representing a wide variety of occupations and organizational hierarchy levels. Participants were on average 27.31 years ($SD$ = 8.41, range 18 to 64). They received either monetary compensation of 15€ or university credit points for completing our study. Voluntariness and anonymity were assured. Study 2 was approved by the Institutional Review Board of the Freie Universität Berlin (No. 88/2014) as part of a funded research project granted to the last author (German research foundation DFG; KR 3457/2-1).

**Study design and materials.** The assessment was conducted in proctored group sessions (up to nine individuals were tested at the same time). All materials were presented to each participant on a computer. The sessions lasted about 90 minutes. Similar to Study 1, a

between-subjects design was used: participants were randomly assigned to one of two conditions: One group received an SJT in its original form (with situation descriptions) and the other group worked on the same SJT, in which situation descriptions were omitted.

Regardless of the condition, every test session followed a three-step procedure: Participants first completed an SJT. Second, a general mental ability test was administered. Third, personality, self-reported teamwork performance, and test motivation were assessed in two randomly administered sequences (to control for fatigue). Peer reports of teamwork performance were either solicited at the end of the test session (if participants had brought a colleague) or collected later on through web links (participants were asked to send these links to their colleagues). Following recommendations by Osborne and Overbay (2004), participants were excluded from further analyses if *z*-scores were below -3 or above +3.

***Situational Judgment Test (SJT).*** The Team Role Test (TRT; Mumford et al., 2008) aims "to measure knowledge of team roles and the contingencies surrounding their appropriate use in team situations" (Mumford et al., 2008, p. 253). For instance, test-takers are asked to decide how to handle team conflicts or how a team works efficiently and productively. The situations depicted a variety of organizational contexts and teams (e.g., sales teams, factory teams, or management teams). In line with prior research (see Krumm et al., 2015), we adapted the SJT to a "pick-the-best" response format. Specifically, three options per item served as distractors (representing ineffective role behavior), whereas one option represented effective role behavior. Test-takers were instructed to select the most effective response option (knowledge-based instruction). The number of correct responses across all TRT items was used for further analyses.

***General mental ability.*** We assessed general mental ability with three subtests (verbal, numerical, figural) of the German version of the General Aptitude Test Battery (GATB; U.S. Employment Service, 1970; German Version: Schmale & Schmidtke, 2001). The administered subtests (spatial aptitude: *Three Dimensional Space* namely a mental folding

task, numerical aptitude: *Arithmetic Reasoning,* and verbal aptitude: *Vocabulary*) were chosen due to their high *g*-loadings (Hunter, 1983). Participants were asked to complete as many items correctly as possible within the time limit specified (6, 7, and 6 minutes, respectively). Reliability (ω) of the general mental ability test scores ranged from .85 to .94 (see Table 5). Correlations among subtests ranged from $r = .25$ to $r = .41$.

***Big Five personality dimensions.*** We measured the Big Five personality dimensions with a short version of the Big Five Inventory (BFI-K; Rammstedt & John, 2005). Participants responded to 21 items on a 5-point rating scale (ranging from 1 = *disagree strongly* to 5 = *agree strongly*). We calculated a mean score per personality dimension. Reliability of the BFI-K ratings ranged from acceptable (.66) to good (.85; see Table 5).

***Criterion measures.*** Self-efficacy for teamwork (eight items; Eby & Dobbins; 1997; e.g., "I can work very effectively in a group setting") and interpersonal adaptability (I-ADAPT; seven items; Ployhart & Bliese, 2006; e.g., "I believe it is important to be flexible in dealing with others") served as specific criteria. Additionally, we included two more general criterion measures: in-role behavior (IRB; seven items; Williams & Anderson, 1991; e.g., "Performs tasks that are expected from him/her") and organizational citizenship behavior (OCBI; seven items; Williams & Anderson, 1991; e.g., "Helps others who have heavy workloads"). All performance ratings were given on a 5-point rating scale.

Criterion data was obtained through self-, peer, and supervisor ratings. Peers had to be colleagues of the target person and had to work on the same team for at least one month (average = 48.07 months). Our results did not differ when duration of working together was controlled for. We obtained peer ratings for 304 participants (between one and four peers per participant). If targets provided ratings from more than one peer, ratings were averaged. Peers (67.9% female) were on average 26.79 years old ($SD = 7.90$, range 18 to 61). Upon completion of the study, we contacted participants again and asked them to contact their current supervisor. We offered an incentive of 30€ for providing a supervisor rating. We

received supervisor ratings for $n$ = 108 participants. As it was key to ensure that the data indeed came from their actual supervisors, we inserted (a) a control question and (b) a question about their self-reported employment relationship. Altogether, 7 (6.0%) supervisor ratings were excluded from the analyses because they failed at least one of these control measures. Supervisors (53.7% female) were on average 42.86 years old ($SD$ = 10.60, range 20 to 74) and had known participants for 46.26 months on average. Reliability estimates for self-, peer-, and supervisor-rated performance criteria ranged from .53 to .89, which is in line with meta-analytic findings on the reliability of job performance ratings (Viswesvaran, Ones, & Schmidt, 1996; see Table 5). Self-, peer, and supervisor ratings of the same performance dimension correlated between $r$ = -.09 and $r$ = .43. Correlations among the different performance components ranged from $r$ = -.09 to $r$ = .48.

***Careless responding.*** Similar to Study 1, we checked for careless responding and inserted the same two bogus items (Levashina et al., 2009). Furthermore, we also checked whether participants who provided supervisor/peer ratings differed from participants for whom we did not receive such ratings. Across both SJT versions, there were no significant differences in terms of SJT scores.

**Results and Discussion**

**Preliminary analyses.** We conducted preliminary analyses to rule out alternative explanations for differences between the SJT versions. First, we verified whether the randomization had worked as expected. Indeed, both groups did not significantly differ in gender, $\chi^2(1) = 1.450$, $p$ = .23, $\phi$ = .05, age, $t(576) = -1.032$, $p$ = .30, $d$ = .09, education level, $\chi^2(8) = 9.629$, $p$ = .29, $\phi$ = .13, or experience in teamwork, $t(576) = -.710$, $p$ = .48, $d$ = .06.

Second, we tested whether scores on the two SJT versions differed in reliability. Reliability estimates of scores on both SJT versions were generally low ($\omega$' s .41 and .45), which is consistent with meta-analyses on the reliability of SJT scores (Campion et al., 2014; Catano et al., 2012; Kasten & Freund, 2016). Importantly, reliability estimates for scores on

the two SJT versions did not differ significantly (with situation descriptions: ω = .41, 95% 95% CI [.33, .50]; without situation descriptions: ω = .45, 95% CI [.36, .54]).

Finally, we tested for metric invariance of SJT scores across the two versions using R Studio (version 1.0.143; R Core Team 2016) and the R package lavaan (version 0.5–22; Rosseel, 2012). As no separate test of metric invariance is available for binary data (Millsap & Yun-Tein, 2004), we followed recommendations by Little et al. (2002) and randomly created parcels of two items each. The overall goodness-of-fit indices for the baseline model (no restrictions) indicated a good fit, $\chi^2(10) = 7.928$, CFI = 1.000, RMSEA = .000 (90% CI [.000, .053]), and SRMR = .024. So, results revealed metric invariance (i.e., factor loadings restricted) across the two SJT versions, $\chi^2(14) = 10.143$ ($\Delta\chi^2(4) = 2.348$, $p = .67$), CFI = 1.000, RMSEA = .000 (90% CI [.000, .042]), SRMR = .029. Note that metric invariance was also achieved when four alternative randomly parceled multi-group models were tested.

**Test of SJT score differences**. We compared overall scores on SJTs with and without situation descriptions. Scores on the SJT with situation descriptions were significantly higher than those on the SJT without situation descriptions, $t(576) = 10.55$, $p < .01$, $d = 0.88$. At the item level (see Table 3), the presence of situation descriptions resulted in significantly higher scores for 7 out of 10 items (with alpha inflation correction this was 6 out of 10). So, the number of the SJT items that could be correctly solved without situation descriptions in Study 2 was less than in Study 1. As a possible explanation, the TRT provides longer, relatively more detailed situation descriptions and therefore may be more context-dependent than the SJTs used in Study 1. In line with McDaniel et al.'s (2016) reasoning, omitting situation descriptions from such an SJT might increase the ambiguity to interpret response options. More generally, these results show that SJT items may best be conceptualized as ranging on a continuum from context-dependent to context-independent (Krumm et al., 2015).

**Test of hypotheses.** When administered with situation descriptions, the SJT showed bivariate correlations with cognitive abilities and broad personality dimensions that were

comparable (albeit somewhat lower) to those previously reported by the SJT developers (Mumford et al., 2008). Thus, the SJT with situation descriptions "behaved" similarly as in prior research. To test our hypotheses about the personality and cognitive ability saturation of SJT scores, we inspected zero-order correlations between the SJT scores and Big Five dimension ratings and cognitive ability test scores, respectively (see Table 6). There were no significant differences in the hypothesized direction (cf. Table 6), lending no support to H1 and H2. We do see a trend for Openness to show a higher correlation with SJT scores in the condition without situation descriptions.

To address Research Question 1 about validity differences between SJT scores with and without situation descriptions, we inspected zero-order correlations with self-, peer-, and supervisor-rated job performance criteria (Table 7[4]). Some exceptions notwithstanding, eyeballing the correlations shows that they were generally higher in the condition with situation descriptions. As we posited a research question and no hypothesis, comparisons of correlations were conducted using two-sided tests. For predicting the broad job performance criteria, there were no significant differences between validities of SJT scores with vs. without situation descriptions. For predicting the specific team-related criteria, SJT scores with situation descriptions were more predictive than the SJT version without situation descriptions. For interpersonal adaptability (peer-rated) and self-efficacy for teamwork (supervisor-rated), the SJT with situation descriptions showed a significantly higher correlation with the criterion than the SJT without situation descriptions.

## General Discussion

Recently, the role of situation descriptions in SJTs has fueled quite some debate. In response to research that situation descriptions are less important for SJT performance than

[4] The average criterion-related validity coefficients for the SJTs with situation descriptions in this study are in line with meta-analytic estimates of SJT validity (McDaniel et al., 2007). Correlations between the SJT and broad performance criteria (peer ratings, $r$ = .15; and supervisor ratings, $r$ = .13) and specific criteria (peer ratings, $r$ = .20, and supervisor ratings, $r$ = .25) fall within the confidence interval reported by McDaniel et al.

typically assumed (Krumm et al., 2015), some scholars have argued for directing more efforts to examining and developing generic and therefore more cost-effective SJTs (e.g., Harvey, 2016). Others posited that it is premature to dismiss situation descriptions in SJTs because this might lower SJT validity and applicant perceptions (e.g., Crook, 2106; Fan et al., 2016; Harris et al., 2016; McDaniel et al., 2016; Melchers & Kleinmann, 2016). To shed light on these unresolved issues, we took a step back and relied on theorizing regarding situation construal to examine whether the absence of situation descriptions impacted on the construct saturation of SJT scores, their validity, and applicant perceptions.

**Implications for Theory**

Our findings provide further evidence that SJT items can be situated on a continuum from tapping into context-dependent to context-independent knowledge: Comparing SJT versions with vs. without situation descriptions results in effect sizes that varied considerably across different SJTs (from $d$ = .16 to .88). In line with Lievens and Motowidlo (2016), one might argue that the personal initiative SJT items rely less on procedural knowledge and more on general domain knowledge. Conversely, the teamwork SJT items used in Study 2 contains a lot of information in the situation descriptions and thus tends to tap more into procedural knowledge and is more situated on the context-dependent side.

The current study is the first to embed responding to SJTs into a situation construal model. This model posits that performance in SJTs with situation descriptions captures the direct effects of person variables (personality and cognitive ability) as well as of situation construal, whereas performance on SJT items without situation descriptions primarily reflects the direct effects of these individual differences (see also Harris et al., 2016; Lievens & Motowidlo, 2016; Sherman et al., 2015; Ziegler & Horstmann, 2017). Across both studies, we found that personality and cognitive ability were generally equally important determinants of SJT performance regardless of condition or SJT type. Hence, these individual differences variables did not emerge as more important drivers of SJT performance when situation

descriptions were absent; they remained just as important as in SJTs with situation descriptions. These results suggest that individual differences such as personality, cognitive ability, and procedural knowledge (and not construal of the situation descriptions) are the main drivers behind SJT performance (Harris et al., 2016; Sherman et al., 2015), lending little support to an interactionist perspective underlying SJT responses. Our findings of the role of personality and cognitive ability fit also well in a context-independent knowledge model of SJT performance (Krumm et al., 2015; Lievens & Motowidlo, 2016; Motowidlo et al., 2006; Motowidlo & Beier, 2010).

The effects of removing situation descriptions on the criterion-related validity of SJT scores differed depending on the breadth of the criteria to be predicted. The validity of SJT scores for predicting broader job performance components (in-role performance and OCB) was generally not affected when situation descriptions were omitted, suggesting situation construal does not play a key role. Conversely, for predicting a more specific criterion such as interpersonal adaptability (peer-rated), SJT scores were more predictive when situation descriptions were presented than when they were omitted, which is in line with the importance of situational construal in adaptive performance models (Pulakos, Arad, Donovan, & Plamondon, 2000). Similarly, SJT scores were more predictive of another specific criterion, namely self-efficacy for teamwork (as rated by supervisors), when situation descriptions were presented. These results suggest that situations might matter when one aims to predict narrow criteria that closely align with the SJT situation descriptions. This might be explained by the fact that point-to-point correspondence between predictor (SJT situation descriptions) and criterion plays a bigger role for more specific criteria than for more global criteria. Future research is needed to explore this explanation further. It was also noteworthy that situations mattered more when the descriptions were longer (see the SJT in Study 2). Apparently, in SJT items with longer situation descriptions, situational construal is important for people to make sense of the items and solve them correctly.

**Implications for Practice**

In terms of practical implications, our study provides some valuable insights for SJT development. Lievens, Peeters, and Schollaert (2008) presented cost estimates for developing SJTs that ranged between $60,00.00 and $120,00.00. The development of situations through involvement of subject matter experts is one of the main drivers of these costs. Our findings and other evidence (Crook et al., 2011; Motowidlo & Beier, 2010; Motowidlo et al., 2006) suggest that there may be less costly but equally valid alternatives. Examples include SJTs in which psychologists formulate more generic situations or single-response SJTs (Motowidlo, Gosh, Mendoza, Buchanan, & Lerma, 2016). Reducing SJT development costs might also further facilitate the proliferation of SJTs in various applied settings.

That said, we do not posit that reduced investment in situation descriptions is the panacea. The attractiveness of SJTs is that they are versatile assessment procedures that exist in various forms and make-ups and can be used for various purposes. For example, in entry-level selection, SJT situation descriptions are typically brief, thereby aiming to predict more general criteria. In such settings, less attention might be paid to developing the situation descriptions, which might move SJTs closer to contextualized personality inventories. However, the opposite might be the case for SJTs in advanced-level selection (e.g., credentialing, selection into advanced training programs), where more elaborate situation descriptions might make sense because in such settings SJTs and their item stems are expected to provide realistic, job-related details in order to be credible and challenging. Neglecting to do this might lower the SJT's power to predict specific criteria. In specific assessment settings, the development of situation descriptions may thus be warranted because even slightly increased validity can be of practical relevance (Cascio & Ramos, 1986; Taylor & Russell, 1939). One might also consider explicitly asking test-takers to make situation judgments in such SJTs. Rockstuhl et al. (2015) found that perception of the situation accounted for substantial variance in SJT performance when the SJT explicitly required

people to make situational judgments. Thus, we generally call for carefully considering how much time and expense to put into the development of situation descriptions compared to other SJT building blocks. In this decision, it is important to balance potential effects on validity and applicant perceptions.

Another practical implication flows from our examination of applicant perceptions. Removing situation descriptions from SJTs had a small and negative effect on two out of six applicant perceptions dimensions. It seems not surprising that stripping off job-related situations affected the candidates' perception of SJTs' job relatedness. We also argued that SJTs without situation descriptions are more difficult and might frustrate test-takers. Yet, little evidence was found. Considering the overall findings on applicant perceptions, the implication is that practitioners should carefully consider applicant reactions in their specific setting, but they should not generally conclude that SJTs with little or no situational content will result in more negative applicant reactions.

Lastly, we also recommend that practitioners more frequently adopt an experimentally-oriented validation approach (Bornstein, 2011; Krumm, Hüffmeier, & Lievens, 2019). This means that one examines which features of a selection procedure are (not) causally related to test scores by manipulating these test features. For example, our results imply that it is not evident to assume that in SJTs elements such as situation descriptions work as intended and, thus, manipulating specific building blocks might be more frequently considered in selection practice and research.

**Limitations**

As a first limitation, our findings were obtained from three different SJTs that tapped into the construct domains of personality and applied social skills. Thus, we did not include an SJT from the third broad construct domain of specific knowledge and skills (Christian et al., 2010). One might suspect that SJTs falling into this domain are more contextualized. However, prior research (Krumm et al., 2015) found that many items of even knowledge-

related SJTs (aviation knowledge) could be solved without the situation description. In addition, only 3% of the currently available SJTs tap into specific knowledge and skills (Christian et al., 2010). So, our results speak to the majority of SJTs.

Second, we included only SJTs with situation descriptions in a written format. However, multimedia and 3D animated formats are becoming increasingly popular (Naemi et al., 2016; Olson-Buchanan & Drasgow, 2006; Weekley & Jones, 1997). Multimedia and 3D animated SJTs show higher face validity (Chan & Schmitt, 1997), improve candidates' attitudes and involvement (Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000), have lower cognitive saturation (Lievens & Sackett, 2006), and outperform text-based SJTs with regard to criterion-related validity for predicting interpersonal criteria (Christian at al., 2010).

Finally, although we argued that stripping off situation descriptions takes out the main context, there might still be situational information in the response options (Harris et al., 2016; Melchers & Kleinmann, 2016) so that participants can try to construe the missing situation description. However, this explanation is not very likely because such contextual information in the response options is typically limited. Moreover, Study 1 revealed negligible differences between the stemless SJT and the situation-neutral SJT. Nevertheless, future research might investigate to what extent participants can construe situation descriptions from response option information.

**Directions for Future Research**

We envision the following avenues for future research. First, we encourage more research on testing moderators of the context-(in)dependency of SJTs. So far, the role of the response options should receive more attention. When the situation description is absent, one possibility is that people might compare the response options to each other or try to construe the missing situation from some context information included in the options (Harris et al., 2016). The availability of trait-related situational cues in SJT situations might be another moderator. Trait activation theory (Tett & Burnett, 2003) posits that variability in behavior

across situations might among others be explained by the situations' relevance for eliciting responses related to a specific trait (Tett & Guterman, 2000). Translating this rationale to SJTs means that SJT situations might (or might not) include situational cues that activate trait-related responses. For instance, a situational cue indicating that a job needs to be done, should prompt test-takers with high Conscientiousness to pick responses that reflect a dutiful and goal-oriented behavior. Conversely, when no trait-relevant situational cues are present in SJT situation descriptions, then no such trait-relevant responses are activated. This might differentially affect the construct saturation of SJT responses.

Second, more research is needed to shed light on the cognitive processes involved when responding to SJT items. Process tracing methods (eye-tracking and verbal protocol analysis) might disentangle the relative importance of situation descriptions and response options. Such research might also show whether and how many times people "cycle back" to the original situational description when choosing among options.

Third, it is also important to consider subgroup differences as an important outcome of personnel selection. Although SJTs show smaller subgroup-differences than general mental ability tests (Bobko & Roth, 2013; Lievens et al., 2006), we do not know whether removing situation descriptions in SJTs affects subgroup differences. To this end, systematic comparisons of diverse samples from different cultures may provide further insights. Relatedly, little is known whether our results hold in different cultures. For instance, one might assume that participants from a culture scoring high on uncertainty avoidance feel more uncomfortable by SJT items without situation descriptions.

Fourth, the selection of situations has received surprisingly little attention in SJT research (Brown et al., 2016). However, various situational taxonomies (e.g., DIAMONDS, Caption; Parrigon, Woo, Tay, & Wang, 2017, Rauthmann et al., 2014) have recently been developed. SJT research has not embraced these developments so far, but they could be a

useful starting point for theory-driven SJT development and studying when and how people's situation construal differs depending on the situation characteristics included (Lievens, 2017).

Finally, research needs to investigate which SJT characteristics engender positive applicant perceptions and when/why contextualization leads to favorable perceptions. Along these lines, an intriguing avenue for future research is to examine how SJTs with brief or virtually no situational details compare to contextualized ("at work") personality inventories.

## Conclusion

SJTs are popular instruments due to their advantages, such as valid prediction of job performance and favorable applicant perceptions. On the other hand, our understanding of how people construe and respond to SJTs is still in its infancy. The current studies add to this growing understanding. We developed a conceptual model of situation construal in SJTs and tested the role of situations and situation construal by omitting the situation descriptions in the item stems. Our results were consistent across the two studies with SJTs from different construct domains. There was little evidence for differences in personality and ability saturation across SJT scores with and without situation descriptions. The effects of omitting situation descriptions on applicant perceptions were also negligible or small. Finally, the impact on criterion-related validity depended on the breadth of the criteria: The validity of SJT scores for predicting global criteria (in-role performance and OCB) was not significantly affected by removing situation descriptions. Conversely, it made a significant difference for predicting specific criteria (interpersonal adaptability, efficacy for teamwork).

**References**

Allport, G. W. (1961). *Pattern and growth in personality*. New York, NY: Hold, Rinehart, and Winston.

Anderson, C. D., Warner, J. L., & Spencer, C. C. (1984). Inflation bias in self-assessment examinations: Implications for valid employee selection. *Journal of Applied Psychology, 69,* 574-580. doi: 10.1037/0021-9010.69.4.574

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695-716. doi: 10.1111/j.1744-6570.1990.tb00679.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9,* 9-30. doi: 10.1111/1468-2389.00160

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology, 54,* 387-419. doi: 10.1111/j.1744-6570.2001.tb00097.x

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12,* 41-75. doi:10.1207/s15328007sem1201_3

Bem, D. J. & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81,* 506-520. doi: 10.1037/h0037130

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative: Towards understanding construct based situational judgment tests. *Personnel Psychology, 62,* 229-258. doi: 10.1111/j.1744-6570.2009.01137.x

Block, J., & Block, J. (1981). Studying situational dimensions: A grand perspective and some limited empiricism. In D. Magnusson (Ed.), *Toward a psychology of situations: An interactional perspective* (pp. 85–106). Hillsdale, NJ: Erlbaum. doi: 10.4324/9780203780886

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

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley. doi: 10.1002/9781118619179

Bornstein, R. F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment, 23,* 532-544. doi: 10.1037/a0022402

Brown, N. A., Jones, A. B., Serfass, D. G., & Sherman, R. A. (2016). Reinvigorating the concept of a situation in situational judgment tests. *Industrial and Organizational Psychology, 9,* 38-42. doi: 10.1017/iop.2015.113

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bruk-Lee, V., Drew, E. N., & Hawkes, B. (2013). Candidate reactions to simulations and media-rich assessments in personnel selection. In M. S. Fetzer & K. A. Tuzinski (Eds.), *Simulations for personnel selection* (pp. 43-60). New York, NY: Springer. doi: 10.1007/978-1-4614-7681-8_3

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York, NY: Springer.

Cabin, R. J., & Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: When and how are the questions. *Bulletin of the Ecological Society of America, 81,* 246-248.

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures: Interactionist psychology operationalized. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 439-456). New York, NY: Routledge. doi: 10.4324/9780203526910.ch19

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

Carroll, S. A., Jones, D. A., & Sulsky, L. M. (2004, April). *Identifying fakers using a bogus item approach.* Presented in the Personnel Selection II Interactive Poster Session at the 19th Annual Meeting of the Society for Industrial and Organizational Psychology, Chicago, IL, USA.

Cascio, W. F., & Ramos, R. A. (1986). Development and application of a new method for assessing job performance in behavioral/economic terms. *Journal of Applied Psychology, 71*, 20-28. doi: 10.1037/0021-9010.71.1.20

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

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

Chen, L., Fan, J., Zheng, L., & Hack, E. (2016). Clearly defined constructs and specific situations are the currency of SJTs. *Industrial and Organizational Psychology, 9*, 34-38. doi: 10.1017/iop.2015.112

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63,* 83-117. doi: 10.1111/j.1744-6570.2009.01163.x

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analyses for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Crook, A. E. (2016). Unintended consequences: Narrowing SJT usage and losing credibility with applicants. *Industrial and Organizational Psychology, 9,* 59-63. doi: 10.1017/iop.2015.118

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

Dahlke, J. A., & Sackett, P. R. (2017). The relationship between cognitive-ability saturation and subgroup mean differences across predictors of job performance. *Journal of Applied Psychology, 102,* 1403-1420. doi:10.1037/apl0000234

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399-412. doi; 10.1111/bjop.12046

Eby, L. T., & Dobbins, G. H. (1997). Collectivistic orientation in teams: An individual and group-level analysis. *Journal of Organizational Behavior, 18,* 275-295. doi: 10.1002/(SICI)1099-1379(199705)18:3%3C275::AID-JOB796%3E3.0.CO;2-C

Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health, 58,* 637-642. doi:10.1007/s00038-012-0416-3

Fan, J., Stuhlman, M., Chen, L., & Weng, Q. (2016). Both general domain knowledge and situation assessment are needed to better understand how SJTs work. *Industrial and Organizational Psychology, 9,* 43-47. doi: 10.1017/iop.2015.114

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175-191. doi: 10.3758/BF03193146

Funder, D. C. (2016). Taking situations seriously: The Situation Construal Model and the Riverside Situational Q-sort. *Current Directions in Psychological Science, 25,* 203-208. doi: 10.1177/0963721416635552

Gatzka, T., & Volmer, J. (2017). Situational Judgment Test für Teamarbeit (SJT-TA) [Situational Judgment Test for Teamwork (SJT-TW)]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. doi: 10.6102/zis249

Göritz, A. S. (2014). Determinants of the starting rate and the completion rate in online panel studies. In Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A. & Lavrakas, P. J. (Eds.), *Online panel research: A data quality perspective* (pp. 154–170). Chichester, England: Wiley. doi: 10.1002/9781118763520.ch1

Göritz, A. S., Borchert, K. & Hirth, M. (2019). Using attention testing to select crowdsourced workers and research participants. *Social Science Computer Review*. Advance online publication. doi:10.1177/0894439319848726

Harris, A. M., Siedor, L. E., Fan, Y., Listyg, B., & Carter, N. T. (2016). In defense of the situation: An interactionist explanation for performance on situational judgment tests. *Industrial and Organizational Psychology, 9,* 23-28. doi: 10.1017/iop.2015.110

Harvey, R. J. (2016). Scoring SJTs for traits and situational effectiveness. *Industrial and Organizational Psychology, 9,* 63-71. doi: 10.1017/iop.2015.119

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

Heidemeier, H. & Göritz, A. S. (2016). The instrumental role of personality traits: Using mixture structural equation modeling to investigate individual differences in the relationships between the Big Five traits and life satisfaction. *Journal of Happiness Studies, 17,* 2595-2612. Doi:10.1007/s10902-015-9708-7

Heydasch, T., Haubrich, J., & Renner, K.-H. (2013). Die Kurzform des Hagener Matrizen-Tests (HMT-S): Ein 6-Item Intelligenztest zum schlussfolgernden Denken [The Short Version of the Hagen Matrices Test (HMT-S): A 6-item test for deductive reasoning.] *Methoden, Daten, Analysen, 7,* 183-208. doi: 10.12758/mda.2013.011

Hogan, R. (2009). Much ado about nothing: The person–situation debate. *Journal of Research in Personality, 43,* 249. doi: 10.1016/j.jrp.2009.01.022

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55. doi: 10.1080/10705519909540118

Hunter, J. E. (1983). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance for the US Employment Services* (Report No. USES-TRR-44). Detroit, MI: Michigan State Department of Labor. (ERIC Document Reproduction Service No. ED236166). doi: 10.1037/e621532009-001

Jackson, D. J., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2017). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology, 90,* 1-27. doi: 10.1111/joop.12151

Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). *European Journal of Psychological Assessment, 32,* 230-240. doi: 10.1027/1015-5759/a000250

Kline, R. B. (2004). *Principles and practices of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Krumm, S., Hüffmeier, J., & Lievens, F. (2019). Experimental test validation: Examining the path from test elements to test performance. *European Journal of Psychological Assessment, 35,* 225-232. doi: 10.1027/1015-5759/a000393

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

Levashina, J., Morgeson, F. P., & Campion, M. A. (2009). They don't do it often, but they do it well: Exploring the relationship between applicant mental abilities and faking. *International Journal of Selection and Assessment, 17*, 271-281. doi: 10.1111/j.1468-2389.2009.00469.x

Lievens, F. (2017). Assessing personality–situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*, 424-440. doi: 10.1002/per.2111

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *Oxford handbook of assessment and selection* (pp. 383-410). New York, NY: Oxford University Press.

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9,* 3-22. doi: 10.1017/iop.2015.71

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

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

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology, 102*, 43-66. doi: 10.1037/apl0000160

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151-173. doi: 10.1207/S15328007SEM0902_1

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills, and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.

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

McDaniel, M. A., List, S. K., & Kepes, S. (2016). The “hot mess” of situational judgment test construct validity and other issues. *Industrial and Organizational Psychology, 9*, 47-51. doi: 10.1017/iop.2015.115

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455. doi: 10.1037/a0028085

Melchers, K. G., & Kleinmann, M. (2016). Why situational judgment is a missing component in the theory of SJTs. *Industrial and Organizational Psychology, 9*, 29-34. doi: 10.1017/iop.2015.111

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36*, 121–140. doi: 10.1177/0149206309349309

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*, 479-515. doi: 10.1207/S15327906MBR3903_4

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252-283. doi: 10.1037/h0035002

Mischel, W. (1977). On the future of personality measurement. *American Psychologist, 32*, 246-254. doi: 10.1037//0003-066x.32.4.246

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

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95*, 321-333. doi: 10.1037/a0017975

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

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance, 29*, 331-346. doi: 10.1080/08959285.2016.1165227

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. doi: 10.1037/0021-9010.91.4.749

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. doi: 10.1037/0021-9010.93.2.250

Naemi, B., Martin-Raugh, M., & Kell, H. (2016). SJTs as measures of general domain knowledge for multimedia formats: Do actions speak louder than words? *Industrial and Organizational Psychology, 9,* 77-83. doi: 10.1017/iop.2015.121

Olson-Buchanan, J. B., & Drasgow, F. (2006). Multimedia situational judgment tests: The medium creates the message. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 253-278). San Francisco, CA: Jossey-Bass.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation, 9,* 1-12.

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2017). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology, 112,* 642-681. doi: 10.1037/pspp0000111

Ployhart, R. E., & Bliese, P. D. (2006). Individual ADAPTability (I-ADAPT) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability. In S. Burke, L. Pierce & E. Salas (Eds.), *Understanding adaptability: A prerequisite for effective performance within complex environments* (pp. 3-39). San Diego, CA: Elsevier. doi: 10.1016/s1479-3601(05)06001-7

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 85,* 612-624. doi: 10.1037/0021-9010.85.4.612

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. [Short version of the Big Five Inventory (BFI-K): Development and validation of an economic inventory for assessment of the five factors of personality]. *Diagnostica*, *51*, 195-206. doi: 10.1026/0012-1924.51.4.195

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., ... Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major

dimensions of situation characteristics. *Journal of Personality and Social Psychology, 107*, 677-718. doi: 10.1037/a0037250

Rauthmann, J. F., Sherman, R. A., Nave, C. S., & Funder, D. C. (2015). Personality-driven situation experience, contact, and construal: How people's personality traits predict characteristics of their situations in daily life. *Journal of Research in Personality, 55*, 98-111. doi: 10.1016/j.jrp.2015.02.003

Reis, H. T. (2008). Reinvigorating the concept of situation in social psychology. *Personality and Social Psychology Review, 12*, 311-329. doi: 10.1177/1088868308321721

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880–887. doi: 10.1037/0021-9010.85.6.880

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

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1-36. doi: 10.18637/jss.v048.i02

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

RStudio Team (2016). RStudio: Integrated Development for R [computer software]. Boston, MA: RStudio, Inc

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2002). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson & H. K. Sinangil (Eds.), *Handbook of industrial, work and organizational psychology* (pp. 165–199). Thousand Oaks, CA: Sage. doi:10.2466/PMS.83.7.1195-1201

Schäpers, P., Lievens F., & Krumm S. (2017, April). *How situational are video-based situational judgment tests?* Presented at the 32th annual convention of the Society for Industrial and Organizational Psychology, Orlando, FL.

Schmale, H., & Schmidtke, H. (2001). *Berufseignungstest BET [Occupational aptitude test BET]*. Bern, Switzerland: Huber.

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

Schmitt, N., & Gilliland, S. W. 1992. Beyond differential prediction: Fairness in selection. In D. M. Saunders (Ed.), *New approaches to employee management: Fairness in employee selection* (Vol. 1, 21– 46). Greenwich, CT: JAI Press.

Serfass, D. G., & Sherman, R. A. (2013). Personality and perceptions of situations from the Thematic Apperception Test. *Journal of Research in Personality, 47*, 708-718. doi: 10.1016/j.jrp.2013.06.007

Sherman, R. A., Nave, C. S., & Funder, D. C. (2013). Situational construal is related to personality and gender. *Journal of Research in Personality, 47*, 1-14. doi: 10.1016/j.jrp.2012.10.008

Sherman, R. A., Rauthmann, J. F., Brown, N. A., Serfass, D. G., & Jones, A. B. (2015). The independent effects of personality and situations on real-time expressions of behavior and emotion. *Journal of Personality and Social Psychology, 109*, 872-888. doi: 10.1037/pspp0000036

Smither, J. W., Reilly R. R., Millsap R. E., Pearlman K., & Stoffey R.W. (1993). Applicant reactions to selection procedures, *Personnel Psychology, 46*, 49-76. doi: 10.1111/j.1744-6570.1993.tb00867.x

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88,* 500-517. doi: 10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34,* 397-423. doi: 10.1006/jrpe.2000.2292

Ullman, J. B., & Bentler, P. M. (2012). Structural equation modeling. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of psychology: Research methods in psychology,* Vol. 2, 2nd ed. (pp. 661–690). Hoboken, NJ: Wiley. doi:10.1002/0471264385.wei0224

United States Department of Labor (1970). *General Aptitude Test Battery: Section III. Development*. Washington, DC: U.S. Government Printing Office.

Vernon, P. A., & Strudensky, S. (1988). Relationships between problem-solving and intelligence. *Intelligence, 12,* 435-453. doi: 10.1016/0160-2896(88)90006-2

Vickers, D., Mayo, T., Heitmann, M., Lee, M. D., & Hughes, P. (2004). Intelligence and individual differences in performance on three types of visually presented optimisation problems. *Personality and Individual Differences, 36,* 1059-1071. doi: 10.1016/s0191-8869(03)00200-9

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557-574. doi: 10.1037/0021-9010.81.5.557

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

Westring, A. J. F., Oswald, F. L., Schmitt, N., Drzakowski, S., Imus, A., Kim, B., & Shivpuri, S. (2009). Estimating trait and situational variance in a situational judgment test. *Human Performance, 22,* 44-63. doi:10.1080/08959280802540999

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

Whetzel, D. L., & Reeder, M. C. (2016). Why some situational judgment tests fail to predict job performance (and others succeed). *Industrial and Organizational Psychology, 9,* 71-77. doi.org/10.1017/iop.2015.120

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17,* 601-617. doi: 10.1177/014920639101700305

Ziegler, M., & Horstmann, K., T. (2017). Importance of testing validity. *European Journal of Personality, 31,* 486-487. doi: 10.1002/per.2119

Table 1

*Sample SJT Item with and without Situation Descriptions in the Item Stem*

| | SJT item with situation description | SJT item without situation description | SJT item without situation description and situation-neutral responses |
|---|---|---|---|
| *Situation description* | A new computer program was installed in your department. No detailed training was provided to save time and money. Some of your colleagues and you feel insecure in dealing with this new program. Errors frequently happen which leads to a loss of time. | *(omitted)* | *(omitted)* |
| *Response instructions* | What would you do? | What would you do? | What would you do? |
| *Response options* | a) I organize an internal training in which more experienced colleagues share their knowledge. (*correct answer*) | a) I organize an internal training in which more experienced colleagues share their knowledge. | a) I organize internal trainings in which more experienced colleagues share their knowledge. |
| | b) I accept working overtime if I have to correct some of the errors | b) I accept working overtime if I have to correct some of the errors | b) I accept working overtime if I have to correct errors. |
| | c) I read books to understand the computer program in my free time to avoid time-consuming errors. | c) I read books to understand the computer program in my free time to avoid time-consuming errors. | c) I read professional books in my free time to avoid time-consuming errors at work. |
| | d) I don't get upset about it because with more practice I will stop making errors. | d) I don't get upset about it because with more practice I will stop making errors. | d) I am not overly worried about errors at work because with more practice I will stop making errors. |

*Note.* Example item taken from an SJT on personal initiative (Bledow & Frese, 2009).

Table 2

*Descriptive Statistics and Reliability Estimates (Study 1)*

| | SJT with situation descriptions | | | | SJT without situation descriptions | | | | SJT without situation descriptions and situation-neutral responses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Measure | *N* | *M* | *SD* | ω | *N* | *M* | *SD* | ω | *N* | *M* | *SD* | ω |
| *Cognitive Ability* | 362 | 2.96 | 1.83 | .71 [.72, .79] | 364 | 3.01 | 1.87 | .76 [.73, .80] | 366 | 2.81 | 1.80 | .74 [.70, .78] |
| *Personality* | | | | | | | | | | | | |
| Extraversion | 104 | 3.26 | 0.91 | .82 [.76, .88] | 100 | 3.09 | 0.99 | .86 [.82, .91] | 96 | 3.11 | 0.92 | .85 [.80, .90] |
| Agreeableness | 104 | 3.17 | 0.80 | .65 [.55, .76] | 100 | 3.09 | 0.69 | .59 [.46, .72] | 96 | 3.11 | 0.79 | .71 [.62, .80] |
| Conscientiousness | 104 | 3.86 | 0.66 | .64 [.52, .76] | 100 | 3.82 | 0.68 | .71 [.61, .81] | 96 | 3.77 | 0.67 | .73 [.64, .82] |
| Neuroticism | 104 | 2.73 | 0.95 | .84 [.79, .89] | 100 | 2.83 | 1.02 | .84 [.78, .89] | 96 | 2.60 | 0.91 | .84 [.78, .89] |
| Openness to experience | 104 | 3.78 | 0.71 | .73 [.65, .81] | 100 | 3.76 | 0.72 | .74 [.66, .82] | 96 | 3.75 | 0.73 | .73 [.64, .82] |
| *SJT 1* | | | | | | | | | | | | |
| Personal Initiative SJT | 362 | 1.06 | 4.15 | .68 [.63, .73] | 364 | .29 | 4.67 | .73 [.68, .77] | 366 | -.22 | 4.28 | .70 [.65, .74] |
| *Applicant Perceptions* | | | | | | | | | | | | |
| Face validity | 362 | 3.77 | 0.64 | .62 [.56, .68] | 364 | 3.66 | 0.66 | .61 [.54, .67] | 366 | 3.65 | 0.67 | .64 [.58, .70] |
| Perceived predictive validity | 362 | 3.11 | 0.76 | .82 [.79, .85] | 364 | 3.05 | 0.79 | .82 [.79, .85] | 366 | 3.12 | 0.76 | .82 [.80, .85] |
| Affect [1] | 362 | 3.70 | 0.88 | .74[.69, .78] | 364 | 3.56 | 0.92 | .69 [.63, .74] | 366 | 3.71 | 0.80 | .68 [.62, .73] |
| Chance to perform | 362 | 3.26 | 0.81 | .83 [.81, .86] | 364 | 3.20 | 0.83 | .83 [.81, .86] | 366 | 3.25 | 0.76 | .81 [.78, .84] |
| Perceived knowledge of results | 362 | 2.90 | 0.88 | .74 [.69, .78] | 364 | 2.88 | 0.79 | .67 [.62, .73] | 366 | 2.98 | 0.79 | .68 [.62, .74] |
| Test-taking motivation | 362 | 4.11 | 0.63 | .87 [.84, .89] | 364 | 4.07 | 0.60 | .81 [.77, .84] | 366 | 4.06 | 0.63 | .85 [.82, .87] |

*Continued on next page*

Table 2 – continued from previous page

| Measure | SJT with situation descriptions | | | | SJT without situation descriptions | | | | SJT without situation descriptions and situation-neutral responses | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *M* | *SD* | ω | *N* | *M* | *SD* | ω | *N* | *M* | *SD* | ω |
| *SJT 2* | | | | | | | | | | | | |
| Teamwork SJT | 281 | 5.69 | 2.73 | .50 [.41, .58] | 270 | 4.21 | 2.78 | .43 [.33, .54] | 288 | 4.00 | 2.50 | .34 [.22, .45] |
| *Applicant Perceptions* | | | | | | | | | | | | |
| Face validity | 281 | 4.14 | 0.61 | .67 [.61, .73] | 270 | 3.88 | 0.69 | .72 [.66, .77] | 288 | 3.89 | 0.65 | .64 [.58, .71] |
| Perceived predictive validity | 281 | 3.22 | 0.76 | .86 [.83, .89] | 270 | 3.14 | 0.75 | .84 [.81, .87] | 288 | 3.16 | 0.78 | .86 [.83, .88] |
| Affect [1] | 281 | 3.95 | 0.75 | .70 [.64, .76] | 270 | 3.79 | 0.83 | .68 [.61, .74] | 288 | 3.80 | 0.82 | .75 [.69, .80] |
| Chance to perform | 281 | 3.39 | 0.75 | .85 [.82, .88] | 270 | 3.25 | 0.84 | .88 [.85, .90] | 288 | 3.22 | 0.77 | .83 [.80, .86] |
| Perceived knowledge of results | 281 | 2.85 | 0.79 | .77 [.72, .82] | 270 | 2.72 | 0.83 | .79 [.74, .83] | 288 | 2.79 | 0.77 | .70 [.64, .76] |
| Test-taking motivation | 281 | 4.23 | 0.58 | .83 [.80, .87] | 270 | 4.17 | 0.60 | .83 [.80, .87] | 288 | 4.14 | 0.60 | .82 [.78, .85] |

*Note*. ω = Omega total; 95% confidence intervals in brackets. Data collection was distributed across three measurement points. Due to participant attrition smaller samples were obtained at measurement point three.

[1] = Affect was measured with a two-item scale, for this reason the Spearman-Brown coefficient was calculated here (see Eisinga, Te Grotenhuis, & Pelzer, 2013

Table 3

*Item-level Effects of the Availability of Situation Descriptions*

| | | With vs without situations | | | | With vs without situations and situation-neutral responses | | | |
|---|---|---|---|---|---|---|---|---|---|
| SJT | Item no. | Cohen's *d* | *t* | *df* | *p* | Cohen's *d* | *t* | *df* | *p* |
| SJT on personal initiative (Study 1) | 1 | 0.37 | 4.934 | 724 | * | 0.81 | 10.955 | 719.10 | * |
| | 2 | -0.13 | -1.694 | 716.41 | | -0.25 | -3.389 | 620.47 | |
| | 3 | 0.17 | 2.325 | 724 | | 0.49 | 6.602 | 726 | * |
| | 4 | -0.09 | -1.203 | 722.34 | | -0.76 | -10.298 | 726 | |
| | 5 | 0.12 | 1.626 | 709.44 | | 0.22 | 2.945 | 725.31 | * |
| | 6 | 0.10 | 1.348 | 718.20 | | 0.41 | 5.511 | 726 | * |
| | 7 | -0.11 | -1.538 | 724 | | -0.13 | -1.741 | 726 | |
| | 8 | -0.23 | -3.117 | 714 | | -0.59 | -8.006 | 708.85 | |
| | 9 | 0.56 | 7.523 | 645.66 | * | 0.78 | 10.491 | 639.01 | * |
| | 10 | 0.08 | 1.042 | 708.73 | | 0.11 | 1.540 | 713.01 | |
| | 11 | -0.07 | -0.937 | 724 | | 0.08 | 1.078 | 726 | |
| | 12 | 0.19 | 2.488 | 696.70 | | 0.34 | 4.636 | 648.95 | * |
| Teamwork SJT (Study 1) | 1 | 0.32 | 3.734 | 526.63 | * | 0.20 | 2.418 | 527.07 | |
| | 2 | 0.19 | 2.249 | 529.90 | | 0.21 | 2.513 | 567 | |
| | 3 | 0.01 | 0.157 | 525.53 | | -0.13 | -1.587 | 547.82 | |
| | 4 | -0.15 | -1.775 | 540.76 | | 0.11 | 1.295 | 567 | |
| | 5 | 0.39 | 4.614 | 548.80 | * | 0.36 | 4.312 | 553.68 | * |
| | 6 | 0.19 | 2.191 | 549 | | 0.12 | 1.411 | 567 | |
| | 7 | 0.65 | 7.617 | 524.95 | * | 0.88 | 10.513 | 567 | * |
| | 8 | -0.12 | -1.392 | 549 | | 0.34 | 4.047 | 557.91 | * |
| | 9 | 0.08 | 0.894 | 537.36 | | 0.13 | 1.578 | 567 | |
| | 10 | 0.29 | 3.374 | 544.38 | * | 0.22 | 2.573 | 567 | |
| | 11 | 0.23 | 2.725 | 549 | * | 0.08 | 0.909 | 567 | |
| | 12 | 0.41 | 4.756 | 524.39 | * | 0.42 | 5.024 | 559.97 | * |
| Team Role Test (Study 2) | 1 | 0.20 | 2.393 | 542.59 | | | | | |
| | 2 | -0.37 | -4.467 | 569.54 | | | | | |
| | 3 | 0.54 | 6.517 | 576 | * | | | | |
| | 4 | 0.12 | 1.444 | 569.83 | | | | | |
| | 5 | 2.20 | 26.685 | 572.75 | * | | | | |
| | 6 | 0.32 | 3.838 | 572.00 | * | | | | |
| | 7 | 0.29 | 3.467 | 576 | * | | | | |
| | 8 | 0.29 | 3.450 | 516.09 | * | | | | |
| | 9 | -0.28 | -3.316 | 496.03 | | | | | |
| | 10 | 0.33 | 3.937 | 498.48 | * | | | | |

*Note*. One-sided *t* tests. Higher effect sizes reflect more correct answers on items with situation descriptions compared with items without situation descriptions. * $p < .005$ ($p$ level adjusted to account for alpha inflation: Study 1: $p$/number of tests = .05/12 = .004; Study 2: $p$/number of tests = .05/10 = .005).

Table 4

*Correlations of SJT Scores with Cognitive Ability and Personality Across Conditions (Study 1)*

| | Bivariate correlation with | | | | | | Difference between correlations (*z*-score) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SJTs with situation descriptions | | SJTs without situation descriptions | | SJTs without situation descriptions and situation-neutral responses | | With vs without situations | | With vs without situations and situation-neutral responses | |
| Measure | SJT PI | SJT TW | SJT PI | SJT TW | SJT PI | SJT TW | SJT PI | SJT TW | SJT PI | SJT TW |
| *Cognitive ability* | .044 | .293** | .125* | .238** | .087 | .177** | -1.095 | 0.691 | -0.580 | 1.459 |
| *Personality* | | | | | | | | | | |
| Extraversion | .093 | .122 | .152 | .155 | -.048 | .131 | -0.421 | -0.214 | 0.983 | -0.057 |
| Agreeableness | .002 | .068 | .039 | -.053 | .138 | .109 | -0.260 | 0.770 | -0.952 | -0.259 |
| Conscientiousness | .117 | .115 | .137 | -.109 | .028 | -.132 | -0.143 | 1.429 | 0.623 | 1.554 |
| Neuroticism | -.169 | -.141 | .057 | -.006 | .126 | .082 | -1.602 | -0.864 | -2.069* | -1.403 |
| Openness to experience | .058 | -.039 | .238* | -.012 | -.085 | .044 | -1.298 | -0.172 | 0.098 | -0.520 |

*Note.* One-sided *z* tests. SJT PI = SJT on Personal Initiative, SJT TW = Teamwork SJT, Cognitive ability: $n_{\text{with situation}}$ = 362 (281), $n_{\text{without situation}}$ = 364 (270), $n_{\text{without situation and situation-neutral responses}}$ = 366 (288); personality questionnaire: $n_{\text{with situation}}$ = 104 (89), $n_{\text{without situation}}$ = 100 (79), $n_{\text{without situation and situation-neutral responses}}$ = 96 (75). Numbers given in brackets refer to the sample sizes with the correlations of the teamwork SJT.
* $p < .05$. ** $p < .01$.

Table 5

*Descriptive Statistics and Reliability Estimates for Study 2 Variables*

| Measure | SJT with situation descriptions | | | | SJT without situation descriptions | | | |
|---|---|---|---|---|---|---|---|---|
| | *N* | *M* | *SD* | ω | *N* | *M* | *SD* | ω |
| *SJT* | | | | | | | | |
| Team Role Test | 307 | 5.88 | 1.53 | .41 [.33, .50] | 271 | 4.48 | 1.65 | .45 [.36, .54] |
| *Cognitive ability* | | | | | | | | |
| Verbal reasoning | 307 | 28.77 | 8.01 | .94 [.93, .95] | 271 | 27.81 | 8.11 | .94 [.93, .95] |
| Numerical reasoning | 307 | 11.26 | 3.20 | .85 [.82, .87] | 271 | 10.80 | 3.34 | .87 [.85, .89] |
| Spatial reasoning | 307 | 21.50 | 5.23 | .89 [.88, .91] | 271 | 20.99 | 5.24 | .89 [.87, .91] |
| *Personality* | | | | | | | | |
| Extraversion | 307 | 3.70 | 0.79 | .83 [.80, .86] | 271 | 3.65 | 0.84 | .85 [.82, .88] |
| Agreeableness | 307 | 3.03 | 0.78 | .70 [.64, .75] | 271 | 2.96 | 0.80 | .72 [.67, .77] |
| Conscientiousness | 307 | 3.78 | 0.65 | .69 [.63, .74] | 271 | 3.72 | 0.63 | .66 [.59, .73] |
| Neuroticism | 307 | 2.90 | 0.94 | .81 [.77, .84] | 271 | 3.04 | 0.85 | .77 [.73, .82] |
| Openness to experience | 307 | 4.13 | 0.69 | .78 [.74, .82] | 271 | 4.13 | 0.62 | .71 [.66, .77] |
| *Job performance* (*self-reports*) | | | | | | | | |
| *Specific (team-related) criteria* | | | | | | | | |
| Self-efficacy for teamwork | 302 | 3.84 | 0.52 | .74 [.70, .79] | 266 | 3.79 | 0.49 | .71 [.66, .77] |
| Interpersonal adaptability measure | 302 | 3.99 | 0.41 | .63 [.57, .69] | 266 | 3.97 | 0.40 | .60 [.53, .67] |
| *Broad performance criteria* | | | | | | | | |
| In-role behavior | 302 | 4.39 | 0.44 | .82 [.78, .85] | 266 | 4.37 | 0.48 | .85 [.82, .88] |
| Organizational citizenship | 302 | 4.01 | 0.43 | .65 [.59, .71] | 266 | 3.98 | 0.37 | .53 [.44, .62] |

*Continued on next page*

Table 5 – continued from previous page

| | SJT with situation descriptions | | | | SJT without situation descriptions | | | |
|---|---|---|---|---|---|---|---|---|
| Measure | *N* | *M* | *SD* | ω | *N* | *M* | *SD* | ω |
| *Job performance (peer-ratings)* | | | | | | | | |
| *Specific (team-related) criteria* | | | | | | | | |
| Self-efficacy for teamwork | 161 | 3.86 | 0.53 | .72 [.65, .79] | 143 | 3.86 | 0.50 | .70 [.62, .77] |
| Interpersonal adaptability measure | 161 | 3.90 | 0.51 | .69 [.62, .77] | 143 | 3.95 | 0.45 | .60 [.50, .70] |
| *Broad performance criteria* | | | | | | | | |
| In-role behavior | 161 | 4.44 | 0.48 | .85 [.81, .88] | 143 | 4.59 | 0.45 | .86 [.82, .89] |
| Organizational citizenship | 161 | 4.09 | 0.46 | .72 [.66, .79] | 143 | 4.13 | 0.48 | .74 [.67, .80] |
| *Job performance* (*supervisor-ratings*) | | | | | | | | |
| *Specific (team-related) criteria* | | | | | | | | |
| Self-efficacy for teamwork | 57 | 3.97 | 0.64 | .82 [.74, .89] | 51 | 3.96 | 0.58 | .74 [.63, .85] |
| Interpersonal adaptability measure | 57 | 4.00 | 0.57 | .82 [.74, .89] | 51 | 4.06 | 0.51 | .77 [.67, .87] |
| *Broad performance criteria* | | | | | | | | |
| In-role behavior | 57 | 4.50 | 0.47 | .76 [.66, .85] | 51 | 4.63 | 0.41 | .89 [.84, .93] |
| Organizational citizenship | 57 | 4.11 | 0.53 | .82 [.74, .89] | 51 | 4.22 | 0.53 | .78 [.68, .87] |

*Note.* ω = Omega total; 95% confidence intervals in brackets.

Table 6

*Correlations of SJT Scores with Cognitive Ability and Personality Across Conditions (Study 2)*

| Measure | Bivariate correlation with | | Difference between correlations (*z*-score) |
|---|---|---|---|
| | SJT with situation descriptions | SJT without situation descriptions | |
| *Cognitive ability* | | | |
| Verbal reasoning | .248** | .263** | -0.192 |
| Numerical reasoning | .119* | .048 | 0.854 |
| Spatial reasoning | .067 | .137* | -0.845 |
| *Personality* | | | |
| Extraversion | .044 (.10) | -.058 | 1.218 |
| Agreeableness | .125* (.16) | .004 | 1.452 |
| Conscientiousness | .118* (.08) | .118 | 0.000 |
| Neuroticism | -.006 (-.10) | .077 | -0.992 |
| Openness | .025 (.16) | .150* | -1.505 |

*Note.* One-sided *z* tests. $n_{\text{with situation}} = 307$, $n_{\text{without situation}} = 271$. Bivariate correlations given in brackets refer to those reported by Mumford et al. (2008) for the original version of the SJT.
* $p < .05$. ** $p < .01$.

Table 7

*Correlations of SJT Scores with Performance Criteria Across Conditions (Study 2)*

| Measure | Bivariate correlation with | | Difference between correlations (*z*-score) |
|---|---|---|---|
| | SJT with situation descriptions | SJT without situation descriptions | |
| *Specific (team-related) criteria* | | | |
| Interpersonal adaptability (*self-rated*) | .105 | .076 | 0.346 |
| Interpersonal adaptability (*peer-rated*) | .195* | -.051 | 2.142* |
| Interpersonal adaptability (*supervisor rated*) | .343* | .113 | 1.223 |
| Self-efficacy for teamwork (*self-rated*) | .047 | .051 | -0.047 |
| Self-efficacy for teamwork (*peer-rated*) | .194* | .009 | 1.615 |
| Self-efficacy for teamwork (*supervisor rated*) | .151 | -.247 | 2.027* |
| *Broad performance criteria* | | | |
| In-role behavior (*self-rated*) | .154** | .140* | 0.169 |
| In-role behavior (*peer-rated*) | .197* | .118 | 0.698 |
| In-role behavior (*supervisor rated*) | .204 | .085 | 0.610 |
| Organizational citizenship behavior (*self-rated*) | .067 | .114 | -0.561 |
| Organizational citizenship behavior (*peer-rated*) | .093 | .056 | 0.321 |
| Organizational citizenship behavior (*supervisor rated*) | .055 | .177 | -0.621 |

*Note.* Two-sided *z* tests. $n_{\text{self-reports; SJT with situation descriptions}} = 302$; $n_{\text{peer ratings; with situation descriptions}} = 161$; $n_{\text{supervisor-ratings; with situation descriptions}} = 57$; $n_{\text{self-reports; SJT without situation descriptions}} = 266$; $n_{\text{peer ratings; SJT without situation descriptions}} = 143$, $n_{\text{supervisor-ratings; without situation descriptions}} = 50$.

* $p < .05$. ** $p < .01$.

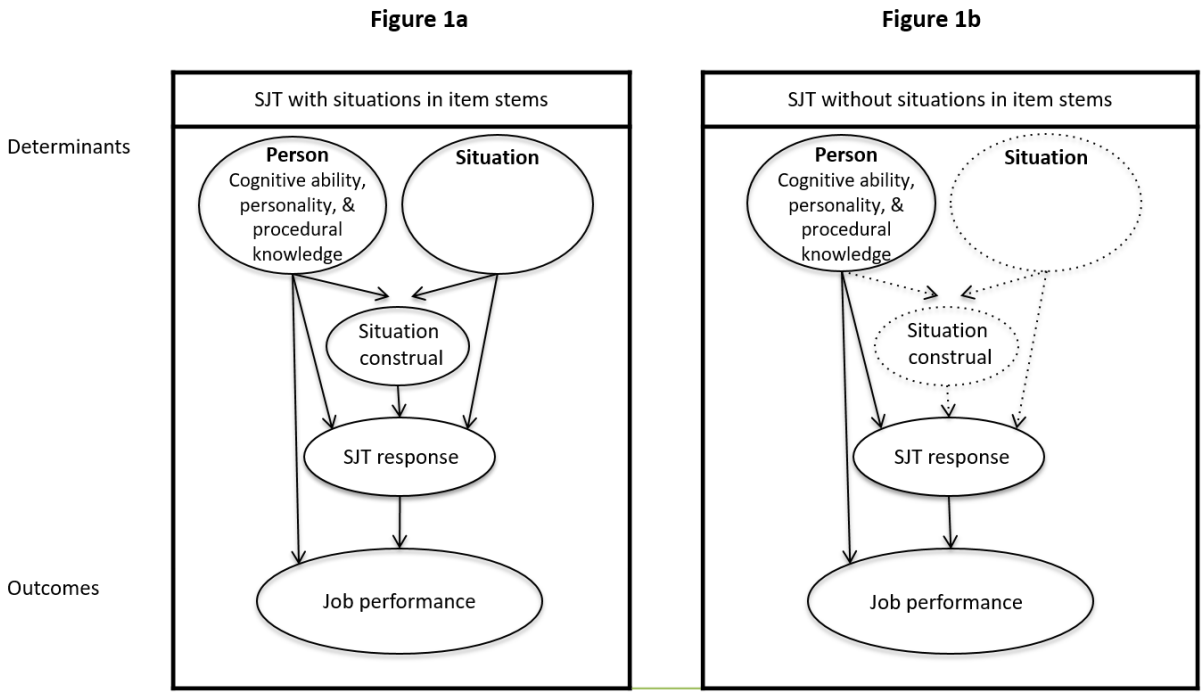


*Figure 1.* Adaptation of the situation construal model (Funder, 2016) to SJTs. Dashed lines mean that the effect of a variable is deemed to be negligible.

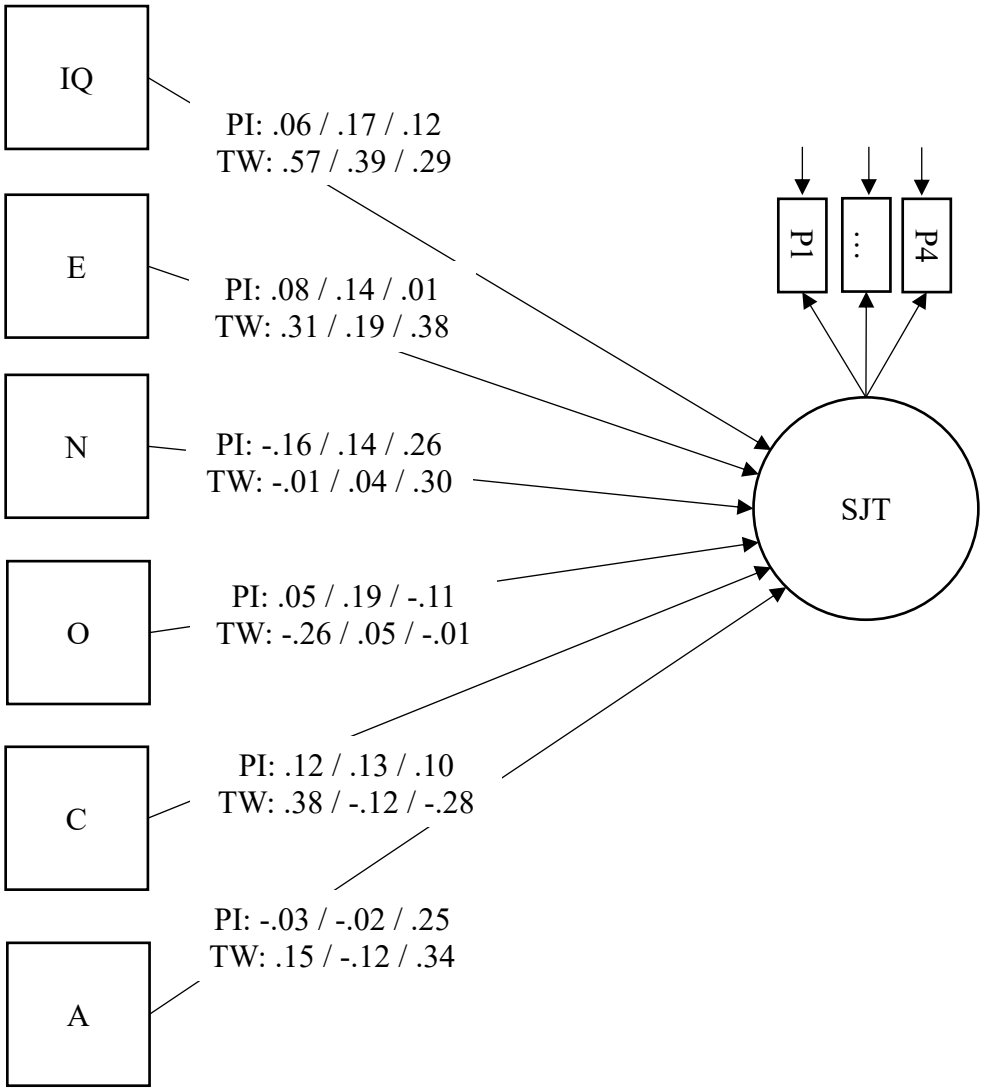


*Figure 2.* Multi-group structural equation model including all constructs (Study 1). Coefficients before dashes refer to the with situation descriptions group, coefficients after first dashes to the without situation descriptions group and after second dashes to the without situation descriptions and with situation-neutral responses group. Direct paths from cognitive ability and personality to criteria as well as covariances the latent cognitive ability and personality variables are omitted for clarity of presentation. Parcels including 3 to 4 items were used as manifest variables. P = parcel; IQ = cognitive ability; E = Extraversion; N = Neuroticism; O = Openness to experience; C = Conscientiousness; A = Agreeableness; TW = teamwork SJT; PI = SJT on personal initiative.