Received: 27 November 2018 | Revised: 24 January 2020 | Accepted: 27 January 2020

DOI: 10.1111/peps.12385

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance

Jan-Philipp Freudenstein[1] | Philipp Schäpers[2] | Lena Roemer[3] | Patrick Mussel[1] | Stefan Krumm[1]

[1]Institute of Psychology, Freie Universität Berlin, Berlin, Germany

[2]Lee Kong Chian School of Business, Singapore Management University, Singapore

[3]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

Correspondence
Jan-Philipp Freudenstein, Institute of Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany.
Email: jpfreudenstein@gmail.com

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: KR 3457/2-1

Abstract

Recent research challenges the importance of situation descriptions for situational judgment test (SJT) performance. This study contributes to resolving the ongoing debate on whether SJTs are situational measures, by incorporating findings on person × situation interactions into SJT research. Specifically, across three studies (N_{Total} = 1,239), we first tested whether situation construal (i.e., the individual perception of situations in SJTs) predicts responses to SJT items. Second, we assessed whether the relevance of situation construal for SJT performance depends on test elements (i.e., situation descriptions and response options) and item features (i.e., description-dependent vs. description-independent SJT items). Lastly, we determined whether situation construal has incremental validity for job-related criteria over and above SJT performance. The results showed that, for most SJT items, situation construal significantly contributed to SJT performance, even if only response options were available. This was also true for SJT items that are significantly more difficult to solve when situation descriptions are omitted (i.e., description-dependent SJT items). Finally, situation construal explained variance in relevant criteria over and above SJT performance. Despite recent efforts to reconceptualize SJTs, our results suggest that they can still be viewed as situational measures. However, situation descriptions may be less crucial for these underlying

situational processes. Theoretical and practical implications are discussed.

KEYWORDS
person × situation interaction, situation construal, situational judgment test, validity

Situational judgment tests (SJTs) are popular instruments in personnel selection, as they exhibit good predictive validity for overall job performance (Christian, Edwards, & Bradley, 2010; McDaniel, Hartman, Whetzel, & Grubb, 2007). When processing typical SJT items, test-takers envision the described situation and pick the response option that reflects how they would most likely behave in such a work situation—at least, this is the predominant understanding of how SJTs work (e.g., Weekley, Hawkes, Guenole, & Ployhart, 2015). In line with this view, SJTs have been traditionally conceptualized as simulations of the relevant work context (Motowidlo, Dunnette, & Carter, 1990). Thereby, situation descriptions were assumed to be the centerpiece of every SJT (e.g., Campion & Ployhart, 2013; Weekley, Ployhart, & Holtz, 2006).

However, several recent studies have revealed inconsistencies in the long-held belief about the importance of situation descriptions (e.g., Jackson, LoPilato, Hughes, Guenole, & Shalfrooshan, 2017; Krumm et al., 2015; Schäpers, Lievens, Freudenstein, Hüffmeier, et al., 2019; Schäpers, Mussel, et al., 2019). For instance, Krumm et al. (2015) showed that, for the majority of items in several different SJTs, it did not make a significant difference whether the situation description was presented or not. The authors concluded, in contrast to previous conceptualizations, that the context in SJTs may be less important for underlying processes. These results led to a debate on how relevant the situation description is for SJTs' functioning (e.g., Crook, 2016; Fan, Stuhlman, Chen, & Weng, 2016; Harris, Siedor, Fan, Listyg, & Carter, 2016; Lievens & Motowidlo, 2016; McDaniel, List, & Kepes, 2016; Melchers & Kleinmann, 2016). In the course of this debate, two opposing views on SJTs emerged. Some scholars agreed with Krumm et al. (2015) that SJTs are less context-dependent than originally assumed (e.g., Crook, 2016; Harvey, 2016; Lievens & Motowidlo, 2016). Other researchers maintained that—even when situation descriptions are taken away—SJTs may still provide relevant context information that test-takers need to understand and interpret. According to the latter view, SJTs can still be conceptualized as context-dependent measures (e.g., Fan et al., 2016; Harris et al., 2016; Melchers & Kleinmann, 2016).

In the current research, we contribute to resolving this controversy by turning our attention to the essence of what constitutes the situation in SJT items: test-takers' psychological construal of the situation (see Brown, Jones, Serfass, & Sherman, 2016). Across three consecutive studies, we incorporate recent findings on person × situation interactions (e.g., Rauthmann et al., 2014). Specifically, we examine to what extent test-takers' psychological construal of a situation affects their responses to SJTs (Study 1). Subsequently, we test whether the relevance of situation construal for SJT performance[1] depends on test elements (i.e., situation descriptions and response options) and item features (i.e., description-dependent vs. description-independent SJT items; Study 2). Finally, we investigate how test-takers' psychological construal of situations has incremental validity over and above SJT performance (Study 3). In doing so, we not only contribute to resolving the ongoing debate on the context dependency of SJTs, but also more generally to a deeper understanding of the situational processes underlying SJT performance. Such an understanding is pivotal for advancing knowledge as to why SJTs work as selection instruments and, from a more practical perspective, how they can be best and cost efficiently developed.

1 | THEORETICAL BACKGROUND

1.1 | Conceptualization of SJTs' underlying processes

SJT items typically consist of work-related situation descriptions and several response options (Weekley & Ployhart, 2006a). Test-takers are usually asked to select the response option that most closely resembles how they would or should behave in the given situation (McDaniel & Nguyen, 2001). Meta-analyses have revealed that SJTs predict overall job performance (Christian et al., 2010), even over and above general mental ability and personality (McDaniel et al., 2007; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Therefore, SJTs enjoy great popularity in applied settings (Lievens, Peeters, & Schollaert, 2008; Ployhart & MacKenzie, 2011; Weekley & Ployhart, 2006b; Whetzel, McDaniel, & Nguyen, 2008).

When reintroducing SJTs to the scientific community, Motowidlo et al. (1990) presented them as low-fidelity job simulations. Similar to assessment center tasks or work samples, SJTs are designed to resemble actual job situations in order to predict on-the-job behavior (Lievens & De Soete, 2012; Motowidlo et al., 1990; Weekley et al., 2015). Consequently, they rest on the assumptions of behavioral consistency and a close resemblance between the simulated content (the situation description in the SJT item stem) and the actual work environment (Bruk-Lee, Drew, & Hawkes, 2013; Lievens & De Soete, 2012; Wernimont & Campbell, 1968). Therefore, situation descriptions in SJT items are often described as the key element for test performance (Campion & Ployhart, 2013; McDaniel & Nguyen, 2001; St-Sauveur, Girouard, & Goyette, 2014; Weekley et al., 2006, 2015; Westring et al., 2009). Accordingly, guidelines for SJT development usually place great emphasis on methods for generating situation descriptions (e.g., the critical incident technique, see Campion, Ployhart, & MacKenzie, 2014; McDaniel & Nguyen, 2001; Motowidlo et al., 1990; Weekley et al., 2006).

In 2015, an experimental study by Krumm et al. (2015) challenged this perspective on SJTs. By omitting situation descriptions from SJT items, they tested whether these descriptions are actually needed to correctly solve SJT items. Surprisingly, the presence or absence of situation descriptions had no influence for between 43% (when *p*-values were not corrected for alpha-inflation) and 71% (when *p*-values were corrected for alpha-inflation) of all items. Krumm et al. (2015) obtained these results for three different SJTs from different construct domains. A further study demonstrated that these results even apply to video-based SJTs (Schäpers, Lievens, Freudenstein, Hüffmeier, et al., 2019). Krumm et al. (2015) argued that test-takers utilize general domain knowledge (i.e., knowledge about generally desirable behavior across a broad range of situations) rather than context-specific knowledge to solve SJT items. This assumption was further corroborated by a recent study that observed only small differences in construct validity and criterion-related validity between SJTs administered with and without situation descriptions (Schäpers, Mussel, et al., 2019). Moreover, these findings are in line with evidence presented by Jackson et al. (2017) who revealed that individual effects rather than situation effects accounted for most of the variance in SJT performance. In addition, Motowidlo and Beier (2010) provided evidence that general beliefs about the effectiveness of trait-related behavior (so-called implicit trait policies, which are unrelated to the situation at hand) predict SJT responses. For instance, some test-takers might believe that agreeable behavior is generally more effective than nonagreeable behavior across a wide range of job-related situations and base their SJT responses upon these beliefs (Lievens & Motowidlo, 2016; see also Motowidlo, Hooper, & Jackson, 2006a, 2006b; Oostrom, Born, Serlie, & van der Molen, 2012). In light of these findings, one might conclude that SJTs are largely context-independent measures (i.e., measures of general domain knowledge).

1.2 | Evidence in favor of the situation

Despite the aforementioned evidence and recent calls for a reconceptualization of SJTs as context-independent measures (Lievens & Motowidlo, 2016), several researchers maintained that situations are in fact relevant to SJTs (e.g., Chen, Fan, Zheng, & Hack, 2016; Fan et al., 2016; Harris et al., 2016; McDaniel et al., 2016; Melchers & Kleinmann,

2016). Rockstuhl, Ang, Ng, Lievens, and Van Dyne (2015) provided empirical evidence for this notion. In their SJT, Rockstuhl et al. (2015) specifically rated participants' evaluations of the presented situations by asking about their thoughts, feelings, and intentions with respect to different people in each situation (i.e., an appraisal of situations). The authors showed that participants' judgments about the presented situation correlated with their reported behavior (i.e., response judgments). However, the results also showed that traditional SJT responses and participants' evaluations of the situation complemented each other in predicting relevant job-related criteria. Notably, Rockstuhl et al. (2015) specifically instructed participants to report their appraisal of the situation. These instructions are typically not given when administering SJTs. Hence, the authors concluded that test developers should put "situational judgment back into SJTs" (Rockstuhl et al., 2015, p. 478).

Another line of research has investigated the relevance of situations in SJTs by disentangling the variance in SJT responses. For instance, Westring et al. (2009) used confirmatory factor analysis to separate variance in SJT responses into trait variance and situational variance. Specifically, they extracted factors capturing interindividual differences across SJT items and factors capturing item-specific variance. They found that situational variance greatly exceeded variance due to individual differences (i.e., trait variance). Similarly, Lievens et al. (2018) made a strong case for the importance of within-person variability in responses across SJT items as a predictor of behavior. They demonstrated that the extent to which test-takers provide inconsistent answers across SJT items can serve as a predictor of performance criteria over and above between-person differences (i.e., SJT scores).

In summary, the results of studies explicitly addressing situation effects on SJT performance are inconsistent. Thus, there is still insufficient empirical evidence to settle the debate about whether SJTs are context-dependent or context-independent measures. In the next section, we argue that a more specific conceptualization and in-depth examination of situations in SJTs is needed to uncover psychologically meaningful effects of situations on SJT performance above and beyond descriptive effects of the context (see Brown et al., 2016).

1.3 | A closer look at situations in SJTs

Like real-life situations, situations in SJT items can be decomposed into three aspects of situational information, namely cues, characteristics, and classes (Brown et al., 2016). *Cues* are physical elements that make up the environmental setting (Rauthmann, Sherman, & Funder, 2015; Saucier, Bel-Bahar, & Fernandez, 2007). As such, they are objective stimuli describing a situation (e.g., a car, a house, a person; Rauthmann et al., 2014). *Characteristics* refer to individuals' psychologically meaningful interpretations of situations (e.g., a situation is stressful; Brown et al., 2016; Fleeson, 2007; Rauthmann et al., 2015). They represent an individual's psychological construal of the situation and encompass the interaction process between situational cues and interindividual variables such as traits, states, and social roles (Fleeson, 2007; Funder, 2016; Mischel & Shoda, 1995; Rauthmann et al., 2014; Reis, 2008; Saucier et al., 2007). Thus, characteristics are individual perceptions of situations and, accordingly, not necessarily identical among individuals (Funder, 2016; Rauthmann, 2015). Lastly, *classes* are aggregate categories of situations including similar cues or characteristics (e.g., work situations; Brown et al., 2016; Rauthmann et al., 2015).

Importantly, it is assumed that behavior is driven by an individual's subjective interpretation of a situation, the situation construal (Funder, 2016; Hogan, 2009; Rauthmann, Sherman, Nave, & Funder, 2015; Reis, 2008).[2] Recently, Brown et al. (2016) argued that this rationale directly translates to situations in SJT items. Given that situations in SJTs are usually only briefly described and thus open to interpretation, these authors suggested that individuals differ in the situation construals they make on the basis of situational cues in SJTs. Furthermore, Brown et al. suggested that individual differences in the perception of situational cues in SJTs (i.e., situation construal) might be pivotal for understanding test-takers' responses to SJT items (see also Mussel, Schäpers, Schulz, Schulze, & Krumm, 2017; Schäpers, Mussel et al., 2019).

The situation construal model was recently incorporated into an empirical study on the underlying processes of SJT performance (Schäpers, Mussel et al., 2019). The authors argued that situation construal is a fundamental

process of SJT responses. Specifically, they "posit that people's differential perceptions of SJT item situations result from the interaction of people's personality and the objective situation" (p. 3). However, they assumed that when situation descriptions are unavailable, situation construal becomes less relevant as an underlying process. Consequently, differences in construct-related validity between SJT versions with and without situation descriptions should emerge. Surprisingly, the authors found no differences in SJT responses' association with personality and cognitive ability between the two SJT versions. They concluded that situation construal may generally be less relevant for the construct-related validity of SJTs. However, in these studies, the assumption that situation construal determines SJT responses was not explicitly tested.

In the current research, we specifically incorporate previous research on situation construal (Funder, 2016; Rauthmann et al., 2015; Reis, 2008; Schäpers, Mussel, et al., 2019). In contrast to previous studies, we directly gauge situation construal for each SJT item. This allows us to explicitly examine the role of situation construal for SJT responses. We argue that SJT performance results from interaction processes between situational cues presented in SJT situation descriptions and response options and interindividual differences (e.g., personality). The results of these interaction processes can be described as perceived situation characteristics (i.e., the test-taker's psychological construal of a situation). In other words, we understand SJT performance as a function of the psychological situation rather than the descriptive context (see also the frame-of-reference effect; Lievens, De Corte, & Schollaert, 2008; Schmit, Ryan, Stierwalt, & Powell, 1995; Shaffer & Postlethwaite, 2012). Accordingly, we generally expect perceived situation characteristics to predict test-takers' responses to SJTs (see Brown et al., 2016; Funder, 2016; Rauthmann et al., 2014; Rauthmann et al., 2015).

***Hypothesis 1*:** Perceived situation characteristics will significantly predict SJT responses.

Although Hypothesis 1 posits that the process of making sense of situational cues in SJTs is relevant for SJT performance, this notion may need further differentiation. That is, elements and features of SJT items may moderate the potential relevance of situation construal for SJT performance. Regarding elements of SJT items, situation construal may be based on either situation descriptions or response options. In fact, several scholars argued that relevant situational cues may still be present when situation descriptions are omitted, that is, when only response options are available (Fan et al., 2016; Harris et al., 2016; Melchers & Kleinmann, 2016). These authors suggested that test-takers may be able to deduce the correct response in SJT items without situation descriptions by closely inspecting the response options and construing the underlying situation from the information they contain (see also Leeds, 2012, 2018). Based on this reasoning, we would expect situation construal to predict SJT performance even when the situation description has been omitted.

However, that may not be the case for all SJT items. Rather, we assume that additional features of SJT items may further moderate this relation. Notably, Krumm et al. (2015) revealed that most, but not all SJT items can be solved without situation descriptions. As such, some SJT items became significantly more difficult when situation descriptions were omitted (Krumm et al., 2015; Schäpers, Lievens, Freudenstein, Hüffmeier, et al., 2019, Schäpers, Mussel, et al., 2019). We hereinafter refer to such SJT items as *description-dependent items* (i.e., item performance decreased in previous research when situation descriptions were omitted) compared to *description-independent items* (i.e., item performance did not decrease in previous research when situation descriptions were omitted). In description-dependent items, the response options may not contain sufficient cues to reconstrue the relevant situations. Thus, perceived situation characteristics may only be meaningful when the situation description is presented, as they cannot be inferred from the response options alone. Conversely, description-independent items may allow for situation construal on the basis of the response options only (Fan et al., 2016; Harris et al., 2016; Melchers & Kleinmann, 2016). Transferring this argument to situation construal, we posit that meaningful situation construal (reflected in a significant prediction of SJT responses) without situation descriptions is possible for description-independent, but not description-dependent SJT items.

***Hypothesis 2*:** Perceived situation characteristics will significantly predict SJT responses, even when situation descriptions are omitted. However, this will only hold for description-independent SJT items.

Two separate processes may contribute to situation construal of SJT items: situational judgment and response judgment. This differentiation was introduced by Rockstuhl et al. (2015). The authors not only asked test-takers about the most effective behavior in a given situation (which they termed response judgment), but they also assessed how test-takers perceive the situation descriptions in these items (which they referred to as situational judgment). Rockstuhl et al. (2015) revealed two results that are of importance for the current study. First, response judgment and situational judgment were correlated yet distinct processes. Second, situational judgment predicted job-related criteria above and beyond response judgment. They concluded that typical SJT scores reflect mostly response judgment and that valuable information remains hidden as situational judgment is typically not captured. In line with Rockstuhl et al. (2015), we argue that the test-takers' perception of situation characteristics for complete SJT items reflects response and situational judgment. Both variance components are particularly relevant for the prediction of job-related criteria. However, SJT scores mostly reflect response judgment. Thus, the situational judgment component in perceived situation characteristics of SJT items will comprise additional variance that predicts job-related criteria.

Our argument rests on the notion that situational judgment more closely resembles situation construal in real-life situations. As delineated above, situation construal is an important psychological driver of behavior (see Hogan, 2009; Mischel & Shoda, 1995; Rauthmann et al., 2014; Reis, 2008). In fact, it has been shown to predict a broad range of real-life behaviors (Parrigon, Woo, Tay, & Wang, 2017; Rauthmann et al., 2014; Sherman, Rauthmann, Brown, Serfass, & Jones, 2015). The same is true for situational judgment and its ability to predict job performance: Like behavior in other domains, job-related behavior stems from an individual's sense-making of specific situations (Jansen et al., 2013; Joseph & Newman, 2010; Tett & Burnett, 2003; Zaccaro, Green, Dubrow, & Kolze, 2018; see also Debusscher, Hofmans, & De Fruyt, 2016; Lievens et al., 2018). This notion was explicitly substantiated by Jansen et al. (2013) who showed that individual differences in situation assessment predicted job performance.

Therefore, we assume that directly assessing perceived situation characteristics for complete SJT items will include the type of judgment that is also relevant in many job-related situations. Hence, we expect perceived situation characteristics of SJT items to explain substantial and unique variance in job-related criteria.

***Hypothesis 3*:** Perceived situation characteristics will significantly predict job-related criteria over and above SJT responses.

2 | OVERVIEW OF STUDIES

In three consecutive studies, we put our working model of SJT performance to the test. As an important incremental contribution, we directly assessed perceived situation characteristics for each SJT item, which has remained a black box in previous studies (e.g., Schäpers, Mussel, et al., 2019). Specifically, we tested our core assumption that perceived situation characteristics of SJT items play a central role in the underlying psychological functioning of SJTs. Thus, we examined whether perceived situation characteristics predict SJT performance (Hypothesis 1; Study 1). Furthermore, we tested whether perceived situation characteristics predict SJT performance even when situation descriptions are absent for both description-dependent and description-independent SJT items (Hypothesis 2; Study 2). Lastly, we examined whether perceived situation characteristics exhibit incremental validity over and above SJT performance (Hypothesis 3; Study 3). All three studies were approved by the Institutional Review Board of the first author's institution (200/2018; "Are Situations just a Relic? The Importance of Situation Characteristics for Situational Judgment Test Performance"). All data and R code are available on the Open Science Framework (osf.io/6kd9h).

3 | STUDY 1

3.1 | Methods

3.1.1 | Participants

A total of 271 individuals took part in Study 1.[3] Participants were recruited in 2017 in Germany via personal contacts, online posts on social media (both job-related and private), and university mailing lists. As an incentive, test-takers were offered feedback on their results on an SJT measuring personal initiative (Bledow & Frese, 2009). In addition, psychology students received course credit. We excluded participants who did not complete at least one full SJT (n = 23). After further exclusion of careless responders (Meade & Craig, 2012; for details see Table S1 in the online supplementary material), a total of N = 227 (156 females, 4 other) participants were included in subsequent statistical analyses. On average, participants' age was M = 24.58 years (SD = 5.52, range 18–66). Almost all participants held a university entrance diploma (95%). Furthermore, 33% of the sample had at least an undergraduate degree and additional 12% had completed vocational education and training (VET; 3 years of vocational training and education for skilled crafts and trades within the German dual system).

3.1.2 | Study design and materials

All data were collected online. Participants responded to items taken from three different SJTs. After each SJT item, participants were asked to indicate the situation characteristics they perceived. To average out possible fatigue effects, all SJTs and all items within each SJT were presented in randomized order.

3.1.3 | Situational judgment tests

Three different SJTs were used. Behavior tendency instructions ("would-do") were applied in all SJTs. That is, we asked participants to indicate what they would most likely do in each situation.

The personal initiative SJT consists of 12 job-related situations with four to five response options each (Bledow & Frese, 2009). We used the original German version of the SJT. Participants' responses were scored as suggested by the test authors, that is, as "1" if they selected the most effective response option, "−1" if they selected the least effective response option, and "0" if they picked one of the other response options. Reliability for this SJT was α = .57 and ω = .57.[4,5] A sample item can be found in the online Supporting Information.

We also administered six items from an SJT measuring self-consciousness (Mussel, Gatzka, & Hewig, 2018). The original test version consists of 22 items in German with four response options each describing everyday public situations with the potential to make someone feel uncomfortable or embarrassed. However, in order to shorten the study duration, we only applied six items. We used Ant Colony Optimization (Leite, Huang, & Marcoulides, 2008; Olaru, Witthöft, & Wilhelm, 2015) to construct a valid short version based on the original validation sample (Mussel et al., 2018; see online Supporting Information for details). For each item, two response options represented high and low trait expressions, respectively. Selecting the option representing high trait expression was scored as "1", all other responses were scored as "−1". Reliability for this SJT was α = .67 and ω = .70. A sample item can be found in the online Supporting Information.

Finally, we used an SJT by Ployhart and Ehrhart (2003) measuring academic achievement and consisting of five critical situation descriptions with four response options each. As this test was only available in English, a native bilingual speaker translated the SJT into German. To check whether this translation produced any inconsistencies or changes in the content, a second bilingual speaker back-translated this SJT to English. We found no differences in content and meaning when comparing the back-translated version to the original SJT. The most effective response option was scored as "1", the least effective response option was scored as "−1", and all other responses were scored as "0". Reliability for this SJT was α = .31 and ω = .34. A sample item can be found in the online Supporting Information.

3.1.4 | Perceived situation characteristics

Rauthmann et al. (2014) developed a taxonomy of perceived situation characteristics. The Situational Eight DIAMONDS describe eight distinct factors, namely Duty (e.g., "Work has to be done"), Intellect (e.g., "Deep thinking is required"), Adversity (e.g., "Somebody is being threatened, accused, or criticized"), Mating, (e.g., "Potential romantic partners are present"), Positivity (e.g., "The situation is pleasant"), Negativity (e.g., "The situation contains negative feelings"), Deception (e.g., "Somebody is being deceived"), and Sociality (e.g., "Social interactions are possible or required"). This taxonomy comprehensively captures psychological representations of situations (Rauthmann & Sherman, 2016a; Rauthmann et al., 2014) and exhibits substantial predictive validity for individual behavior over and above personality (Parrigon et al., 2017; Rauthmann & Sherman, 2016a, 2016b; Rauthmann et al., 2014, 2015).

To assess the individually perceived situation characteristics of the SJT items, participants filled out either the S8* (Rauthmann & Sherman, 2016a) or the S8-I (Rauthmann & Sherman, 2016b) on a 7-point rating scale (1 = *does not apply at all*; 7 = *applies completely*) after each SJT item. Both measures capture the Situational Eight DIAMONDS, with the S8* consisting of three items for each of the eight facets and the S8-I consisting of one item for each facet. All items in the German versions of the S8* and S8-I were pilot tested and, if necessary, rephrased slightly to avoid ambiguity.

Participants were randomly assigned to fill out the S8* for one of the three SJTs (n_{PI} = 82; n_{SC} = 72; n_{P} = 73). To shorten the study duration, the S8-I was presented for the remaining two SJTs. Responses for perceived situation characteristics were collected for all 23 SJT items. The reliability coefficients for the three S8* items measuring each of the DIAMONDS dimensions, averaged across all 23 SJT items, were α = .66 (*SD* = .08) for Duty, α = .73 (*SD* = .06) for Intellect, α = .71 (*SD* = .11) for Adversity, α = .57 (*SD* = .11) for Mating, α = .71 (*SD* = .10) for Positivity, α = .80 (*SD* = .44) for Negativity, α = .71 (*SD* = .09) for Deception, and α = .60 (*SD* = .14) for Sociality. Albeit somewhat low, internal consistencies are overall in line with coefficient alpha values reported by Rauthmann and Sherman (2016a, range: .61–.90). The S8* items were aggregated to form a mean score for each facet. See Table 1 for pooled correlations for each DIAMONDS dimension across all SJT items.

3.1.5 | Data analyses

As SJTs are usually not designed to ensure the test items' homogeneity in terms of perceived situation characteristics, we did not expect items within the same SJT to elicit a homogeneous set of perceived situation characteristics. Therefore, our analyses focused on individual items rather than the aggregated SJT test scores. To estimate the overall effect of perceived situation characteristics on SJT performance across all SJT items, we utilized mixed-effect models for ordered dependent variables with crossed random effects for SJT items and subjects (Baayen, Davidson, & Bates, 2008; Tutz & Hennevogl, 1996). This procedure makes it possible to assess the overall relation between perceived situation characteristics and SJT item performance (fixed effects) and to simultaneously account for unique variance in SJT performance (random intercepts) and perceived situation characteristics (random slopes) due to subjects and SJT items (Baayen et al., 2008; Tutz & Hennevogl, 1996). Specifically, the Situational Eight DIAMONDS served as fixed predictors of SJT item responses. We further allowed different regressions weights for perceived situation characteristics within each SJT item (random slopes). We centered perceived situation characteristics within persons and further included the grand mean-centered average of each of the DIAMONDS dimensions across all SJT items as a predictor on the person level (Enders & Tofighi, 2007; see also Sherman et al., 2015). This Level 2 predictor controls for general person effects due neither to situations nor to person × situation interactions (i.e., the tendency to perceive all SJT items in the same manner, independent of the specific situation). The significance of effects was evaluated with likelihood ratio tests and the Horowitz adjustment of McFadden's pseudo-$R^2_{McF/H}$ (Hemmert, Schons, Wieseke, & Schimmelpfennig, 2016; Horowitz, 1982). Hox (2010) suggests that random effects models adequately deal with missing data as they incorporate full information into the analysis (see also Hedeker & Gibbons, 1997; Snijders, 1996). For additional information, see Table S1.

TABLE 1 Pooled descriptive statistics of the DIAMONDS across SJT items

	M (*SD*)	1	2	3	4	5	6	7	8
Study 1									
1. Duty	4.60 (1.40)	–							
2. Intellect	4.00 (1.61)	.46	–						
3. Adversity	2.30 (1.34)	.07	.08	–					
4. Mating	1.72 (1.19)	.00	.07	.17	–				
5. Positivity	2.54 (1.27)	−.03	.06	−.09	.16	–			
6. Negativity	4.27 (1.50)	.11	.11	.35	.01	−.31	–		
7. Deception	1.89 (1.20)	.06	.10	.30	.22	.05	.20	–	
8. Sociality	4.05 (1.76)	.09	.18	.11	.24	.23	.07	.11	–
Study 2									
1. Duty	5.04 (1.72)	–							
2. Intellect	4.51 (1.69)	.45	–						
3. Adversity	2.08 (1.44)	.03	.11	–					
4. Mating	1.73 (1.28)	−.03	.03	.31	–				
5. Positivity	2.94 (1.41)	.03	.06	−.01	.17	–			
6. Negativity	4.02 (1.60)	.07	.11	.26	.08	−.31	–		
7. Deception	2.22 (1.44)	−.01	.08	.42	.28	−.03	.27	–	
8. Sociality	4.11 (1.79)	.09	.13	.09	.21	.23	.06	.08	–
9. Group 1	–	−.01	−.02	−.10	−.04	−.01	−.08	−.08	.05
10. Group 2	–	.08	.04	.18	.05	.18	−.07	.04	.03
11. Group 3	–	−.06	−.01	−.07	−.01	−.16	.14	.04	−.07
Study 3									
1. Duty	4.29 (1.48)	–							
2. Intellect	3.82 (1.68)	.40	–						
3. Adversity	2.63 (1.51)	.11	.20	–					
4. Mating	1.83 (1.22)	.03	.12	.15	–				
5. Positivity	2.41 (1.24)	.01	.07	−.10	.18	–			
6. Negativity	4.32 (1.58)	.09	.10	.34	.04	−.32	–		
7. Deception	2.07 (1.44)	.06	.19	.34	.20	−.02	.25	–	
8. Sociality	4.23 (1.83)	.13	.24	.11	.24	.23	.03	.17	–

Note. Sample sizes in Study 1 ranged between $n = 209$ and 224. Sample sizes in Study 2 ranged between $n = 561$ and 632; $n_{group1} = 261$, $n_{group2} = 214$, $n_{group3} = 252$. Sample sizes in Study 3 ranged between $n = 284$ and 285.

3.2 | Results

3.2.1 | Preliminary analysis

First, we checked whether participants' perceived situation characteristics differed across SJT items. A repeated measure MANOVA for the eight DIAMONDS across all SJT items revealed a significant main effect, $F(22, 4952) = 64.40$, $p < .001$, $\eta^2 = .22$. The effect was also present for all DIAMONDS when conducting separate ANOVAS. Therefore, the results suggest that perceived situation characteristics differed across the 23 SJT items.

We also applied generalizability theory analysis (Brennan, 2001; Shavelson, Webb, & Rowley, 1989) to determine the amount of reliable variance in the DIAMONDS that can be attributed to either persons (i.e., similar ratings across

SJT items) or SJT items (i.e., situation-specific ratings). On average, 31.4% (SD = 15.3) of the variance in perceived situation characteristics could be attributed to differences among SJT items. However, 21.3% (SD = 9.1) of the variance could be attributed to persons. This indicates that individuals have a certain general tendency to evaluate perceived situation characteristics similarly across SJT items. These findings justify our approach of controlling for overall person effects (in perceived situation characteristics) when examining the relevance of perceived situation characteristics for SJT performance.

3.2.2 | Hypothesis tests

We applied mixed-model regressions to test the relations between perceived situation characteristics and SJT performance while controlling for the dependency among subjects and different SJT items (i.e., random intercepts). Compared to the null model (i.e., fixed intercept only), including a random intercept for SJT items significantly increased model fit, $\Delta\chi^2(1) = 1{,}566.10$, $p < .001$, $R^2_{\text{McF/H}} = .143$. The same was true for the random intercept for subjects, but only if adjusted for the SJT items' nested structure within three different SJTs, $\Delta\chi^2(6) = 554.19$, $p < .001$, $R^2_{\text{McF/H}} = .050$. Thus, effects due to items and individuals accounted for reliable variance in SJT responses. Notably, the effect due to SJT items exceeded the effect due to individuals.

For the perceived situation characteristics Adversity, Positivity, Negativity, and Deception, significant fixed effects were found, thus indicating their overall importance for SJT performance (see Table 2). Furthermore, for six of the eight DIAMONDS (with Mating and Deception being the exceptions), the random slope accounted for a significant amount of variance in SJT performance. The significant random slopes demonstrate the heterogeneity of perceived situation characteristics relevant for SJT performance across items (i.e., which perceived situation characteristics predict SJT responses differs across SJT items). The effects were also present when corrected for individuals' general tendency to perceive situations in a certain way (grand mean-centered averages of perceived situation characteristics), even though the average of Mating and Positivity across all SJT items substantially predicted responses to SJT items as well. Overall, including perceived situation characteristics significantly improved model fit compared to a model with only random intercepts and the grand mean-centered averages of perceived situation characteristics, $\Delta\chi^2(52) = 890.32$, $p < .001$, $R^2_{\text{McF/H}} = .090$. In sum, these results lend support to Hypothesis 1.

3.3 | Discussion

Study 1 provided evidence supporting the assumption that perceived situation characteristics can explain responses to SJTs: All DIAMONDS (with the exception of Mating) significantly predicted performance on SJT items. Notably, we found that the facets Adversity, Positivity, Negativity, and Deception predicted SJT performance across all SJT items. Thus, our findings lend support to the situation-dependent perspective on SJTs. That is, situation construal seems to matter for SJT performance (cf., Schäpers, Mussel et al., 2019). This is further corroborated by the finding that the proportion of SJT item variance accounted for by person main effects was smaller than the proportion of SJT item variance accounted for by situation-specific effects (see Westring et al., 2009; cf., Jackson et al., 2017).

Study 1 also revealed that the relevance of different facets of perceived situation characteristics as well as the general importance of situation construal differed across items. In other words, some SJT items were more dependent on situation construal than others. Such differences in the relevance of situation construal may explain why, in previous studies, some but not all SJT items could still be solved when situation descriptions were omitted (Krumm et al., 2015; Schäpers, Lievens, Freudenstein, Hüffmeier, et al., 2019). Study 2 will specifically examine whether the relevance of situation construal for SJT item performance depends on test elements (i.e., situation descriptions vs. response options) and item features (i.e., whether it differs between description-dependent vs. description-independent SJT items).

TABLE 2 Mixed-model results (Study 1)

	Fixed effects				Random effects		Person level		Correlations among random effects							
	B	*SE*	*p*	OR	σ^2	*p*	B_{Mean}	*p*	Item	D	I	A	M	O	N	De
Intercept (item)					2.75											
D	0.07	0.06	.290	1.07	0.07*	<.001	0.05	.459	−.08							
I	0.09	0.06	.126	1.09	0.05*	<.001	0.07	.261	.18	−.16						
A	0.15*	0.05	.002	1.16	0.03*	<.001	0.08	.204	.24	−.56	−.26					
M	−0.05	0.05	.303	0.95	0.01	.663	−0.12*	.030	−.11	.52	−.62	−.05				
O	−0.32*	0.09	.001	0.72	0.17*	<.001	−0.22*	.003	−.38	−.04	.37	−.70	−.51			
N	0.14*	0.07	.040	1.15	0.08*	<.001	0.09	.085	.20	−.34	−.34	.66	.37	−.81		
De	−0.11*	0.05	.011	0.89	0.02	.435	−0.06	.397	.15	.30	.63	−.81	−.40	.75	−.79	
S	0.11	0.06	.052	1.11	0.05*	<.001	0.05	.191	−.21	−.07	.01	−.27	.45	.18	−.07	.10
Log-likelihood −4,189.89																

Note. Individual responses for SJT items served as dependent variable. Random effects refer to a random intercept for SJT items and random slopes on item level for the DIAMONDS. A random intercept for individuals was also included (not depicted) to account for the nested structure within individuals. B_{Mean} refers to the grand mean-centered average of the DIAMONDS for all SJT items as person level predictor. $N = 227$.

Abbreviations: A, Adversity; D, Duty; De, Deception; I, Intellect; M, Mating; N, Negativity; O, Positivity; S, Sociality.

*$p < .05$; *p*-values for random effects refer to likelihood ratio tests between models with and without the corresponding parameter.

4 | STUDY 2

To test Hypothesis 2, we deployed a between-subjects experimental design that aimed at separating the unique influence of situation descriptions and response options on the relevance of situation construal for SJT performance. Group 1 received the entire SJT item; thus, both situation descriptions and response options were potential sources of situation construal for these test-takers. For Group 2, we omitted the situation descriptions. Hence, this group was only able to base their situation construal on the response options. Finally, Group 3 saw each SJT item's situation description without response options and then completed the situation construal questionnaire. Only after that did they receive the response options for the SJT item (because we wanted to gauge their SJT item performance as well). In other words, this group made their situation construals based on situation descriptions only. Additionally, we differentiated between items where situation descriptions had high and low relevance for item performance (i.e., description-dependent and description-independent SJT items; this distinction was determined a priori based on prior studies). Thus, this study sheds light on the comparative relevance of psychological situation construal on the basis of different item elements and features for SJT performance (Hypothesis 2).

4.1 | Methods

4.1.1 | Participants

We conducted an a priori power analysis by applying Monte-Carlo simulation to determine the sample size required to detect effects similar to those found in Study 1 (see Muthén & Muthén, 2002). As Hypothesis 2 partly specifies nonsignificant effects, it is appropriate to define α and β equally. A total sample of 618 participants (206 per group) was needed to ensure sufficient power ($1 - \beta = .95$) with $\alpha = .05$. Overall, 791 individuals were recruited in 2017 in Germany via personal contacts (e.g., e-mail), postings in online career communities, and social media. As an incentive, test-takers were offered feedback on their Big Five personality dimensions. In addition, psychology students received course credit. We excluded participants who did not complete at least one SJT item along with the corresponding items assessing perceived situation characteristics ($n = 14$). After additional exclusion of careless responders, $N = 727$ (324 females, 2 other; age: $M = 30.74$, $SD = 11.26$, range: 18–70)[6] were included in the statistical analyses. On average, test-takers reported $M = 7.10$ ($SD = 10.58$) years of work experience and $M = 30.15$ ($SD = 16.70$) average weekly working hours. In total, 72% held at least an undergraduate degree. Although participants worked in a broad range of industries (e.g., banking, manufacturing, IT), the most commonly indicated fields of employment were academia (24%) and the pharmaceutical industry (18%).

4.1.2 | Study design and materials

All data were collected online in a between-subjects design with participants randomly assigned to three groups ($n_{group1} = 261$; $n_{group2} = 214$; $n_{group3} = 252$). Participants responded to a total of 12 items taken from three different SJTs. For each SJT, we chose two items for which previous research had found no mean differences in item performance when presented with and without situation descriptions (i.e., description-independent items). We chose another two items from each SJT for which previous research had found large differences when administered with versus without situation descriptions (i.e., description-dependent items; for example, see online Supporting Information). After each SJT item, participants were asked to report their perceived situation characteristics.

Situational judgment tests

We applied four items from a German SJT measuring knowledge about functions on Facebook (e.g., privacy settings, Messenger; Schäpers, Lievens, Freudenstein, Schulze, et al., 2019). All items describe situations related to using Facebook and require knowledge of the functionality of several Facebook settings. We asked participants to choose the most effective behavior among four response options. Responses were scored as "1" if participants selected the most

effective behavior. All other responses were scored as "0". We chose the two most description-dependent and the two most description-independent items based on previous results by the SJT's authors. A sample item can be found in the online Supporting Information.

In addition, we applied four items from the Team Role Test (TRT; Mumford, Van Iddekinge, Morgeson, & Campion, 2008). This widely used SJT assesses team role knowledge. Again, we chose two description-dependent and two description-independent items from a modified and translated to German version by Schäpers, Mussel, et al. (2019). This version asks participants to pick the most effective response among four options. Thus, selecting the most effective response option was scored as "1"; all other responses were scored as "0". A sample item can be found in the online Supporting Information.

We also applied four items from the personal initiative SJT (for details, see Study 1; Bledow & Frese, 2009). We selected the two most description-dependent and description-independent items based on previous findings by Schäpers, Mussel, et al. (2019).

Perceived situation characteristics

Similar to Study 1, we applied the S8-I (Rauthmann & Sherman, 2016b) to assess each individual's perceived situation characteristics for every SJT item. As the S8-I consists of only one item per facet and no complete SJTs were used, no reliabilities are reported.

Further measures

We asked participants about their level of experience with the different SJT domains using single-item indicators ("How often do you use Facebook?" 1 = *monthly or infrequently* to 5 = *several times a day*; "How much work experience do you have in teams?" 1 = *no experience*, 5 = *plenty of experience*; "How proactive are you in a work context?" 1 = *not proactive*, 5 = *very proactive*). We further applied the BFI-2-XS (Rammstedt, Danner, Soto, & John, 2018) as a control measure of group differences. This test consists of 15 items assessing Big Five personality on a 5-point rating scale (1 = *disagree strongly* to 5 = *agree strongly*). Cronbach's alphas ranged from $\alpha = .41–.71$.

4.1.3 | Data analyses

The results of Study 1 demonstrated that the relevance of perceived situation characteristics for SJT responses varied considerably across SJT items. Thus, our analyses focused on the item level. We conducted multigroup regression analyses for each SJT item. All participants who completed the SJT item of interest and the corresponding assessment of perceived situation characteristics were included in the analysis. In a preliminary step, we computed baseline models for which the SJT item response served as the dependent variable and the residualized DIAMONDS as eight predictor variables. Residual scores were calculated by regressing the DIAMONDS on the grand mean-centered averages of the DIAMONDS across SJT items. This was done to control for the general tendencies in individuals' perceived situation characteristics and to retain model simplicity (Wurm & Fisicaro, 2014). Next, all coefficients were freely estimated for all three groups. Afterwards, we constrained all regression coefficients across groups to equality and tested this model against the baseline model via scaled χ^2-difference tests (Satorra, 2000). If this constrained model had significantly worse fit, we compared regression weights between two groups only in a stepwise manner (i.e., comparison of regression weights between Groups 1 and 2, Groups 1 and 3, and Groups 2 and 3). Overall, model fit was evaluated based on scaled χ^2-difference tests against the null model and R^2. For additional information, see Table S2.

4.2 | Results

4.2.1 | Preliminary analyses

To rule out between-group effects due to sampling error, we tested for group differences in demographic variables and personality. The groups did not differ in gender ratio, $\chi^2(4) = 1.019$, $p = .963$, Cramer's $V = .03$; age, $F(2, 539) = .47$,

$p = .624$, $\eta^2 = .00$; educational level, $\chi^2(10) = 8.513$, $p = .579$, Cramer's $V = .09$; work experience, $F(2, 537) = .28$, $p = .754$, $\eta^2 = .00$; or weekly working hours, $F(2, 515) = 1.40$, $p = .247$, $\eta^2 = .01$. Furthermore, the groups did not differ in Big Five personality, $F(2, 539) = 1.70$, $p = .087$, $\eta^2 = .02$. Finally, no group differences were found in self-reported frequency of Facebook use, $F(2, 470) = 2.86$, $p = .058$, $\eta^2 = .01$, self-reported frequency of teamwork, $F(2, 539) = .14$, $p = .867$, $\eta^2 = .00$, or self-reported initiative in work contexts (single-item measure), $F(2, 539) = 1.30$, $p = .272$, $\eta^2 = .00$.

To test whether all SJT items fell into the expected category of description (in)dependency, we applied one-sided *t*-tests for SJT item performance between Groups 1 and 2 (see Krumm et al., 2015). Contrary to our assumptions, mean differences were found for two items where we did not expect any difference, while no difference was found for one item where a difference was expected (see Table S4). Therefore, we removed these three items from subsequent analyses. Notably, the interpretation of the main results presented below did not differ when recategorizing the three items (see Table S5).

In Group 3, perceived situation characteristics were assessed after presenting the situation description without response options. The response options only became visible after the perceived situation characteristics were assessed, which might have altered participants' responses. However, no differences in item difficulty were found between Groups 1 and 3, thus indicating that assessing perceived situation characteristics in between seeing the situation descriptions and responding to the SJT item had no influence on test performance.

We also tested whether our experimental manipulation affected the assessment of perceived situation characteristics for each SJT item. As only one item was available for each DIAMONDS dimension, we compared the pooled correlations among the DIAMONDS across all SJT items. The comparison revealed no significant differences, $\chi^2{}_{g1g2}(56) = 17.51$, $p = .99$; $\chi^2{}_{g1g3}(56) = 12.98$, $p = .99$; $\chi^2{}_{g2g3}(56) = 26.10$, $p = .99$ (see Table 1 for pooled correlations among DIAMONDS across SJT items).

Finally, we tested whether the DIAMONDS differed across SJT items and groups. MANOVA results indicated significant main effects for group membership, $F(2, 7013) = 41.5$, $p < .001$, $\eta^2 = .05$, and SJT item, $F(11, 7013) = 83.17$, $p < .001$, $\eta^2 = .12$, as well as a significant interaction effect, $F(22, 7013) = 9.34$, $p < .001$, $\eta^2 = .03$. Separate ANOVAs revealed that these effects were equally present for all DIAMONDS. Graphical inspection of the interaction effect confirmed the heterogeneous mean differences in perceived situation characteristics across groups and across SJT items (see Figure S6).

4.2.2 | Hypothesis tests

Overall, in all three groups, at least one dimension of DIAMONDS significantly predicted performance for eight of nine SJT items. For one description-dependent SJT item from the personal initiative SJT, DIAMONDS predicted SJT performance only in Groups 1 and 2 (see Table 3). However, for two description-dependent items the overall model fit of the baseline model did not differ significantly from zero, even though DIAMONDS significantly predicted SJT performance. That is, the effect sizes for DIAMONDS predicting SJT responses on these items were relatively small (mean $|\beta| = .26$, $SD = 0.06$). Nevertheless, when the alpha level was corrected for the number of predictors (Bonferroni correction; $p = .05/8 = .00625$; Cabin & Mitchell, 2000) perceived situation characteristics still significantly predicted SJT item responses for one of those two items.

Results for description-independent items

In the next step, we constrained all regression weights across all groups to equality and tested whether the restricted model differed significantly from the freely estimated model (baseline model). Hence, we tested whether the relevance of perceived situation characteristics for SJT performance differed across groups. For all description-independent items, the restricted model did not differ from the baseline model for Groups 1 and 2. Moreover, for one of the four items, the relevance of situation construal for SJT performance did not differ in Group 3 either (see Table 3). Thus, the results partly support Hypothesis 2, as DIAMONDS equally predicted SJT performance in description-independent items regardless of the presence or absence of a situation description.

TABLE 3 Multigroup regression analysis (Study 2)

	Comparison against null model			Comparison against baseline model			Relevant DIAMONDS			R^2		
	χ^2	*p*	Equality constraints	$\Delta\chi^2$	Δdf	*p*	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
Description-independent items												
TRT 4	61.235	<.001	All groups	12.145	6.39	.072	D, De	D, De	D, De	.20	.16	.21
PI 7	88.013	<.001	Groups 1 and 2	5.767	3.92	.209	I, De, S	I, De, S	S	.25	.22	.14
PI 10	55.987	<.001	Groups 1 and 2	8.762	3.60	.052	D, I, S	D, I, S	O, N	.18	.20	.08
FB 8	51.704	<.001	Groups 1 and 2	5.061	2.96	.163	D, M	D, M	O, De	.14	.19	.12
Description-dependent items												
TRT 3	50.846	.001	All groups	11.711	5.80	.062	D, S	D, S	D, S	.14	.13	.13
TRT 5	47.299	.003	All groups	5.664	5.58	.409	I	I	I	.28	.33	.23
PI 1	75.476	<.001	–	–	–	–	D, A, O, S	I, A, S	–	.26	.34	.04
PI 9	28.693	.232	All groups	5.198	5.93	.510	D, N	D, N	D, N	.08	.10	.08
FB 1	35.379	.063	All groups	7.411	6.12	.296	S	S	S	.05	.07	.05

Note. Comparison against null model refers to the comparison of the model without equality constraints to zero. All χ^2 values of this comparison with $df = 24$. All columns to the right of the comparison against the null model refer to the model with equality constraints. Comparison against baseline model refers to comparison of the model with equality constraints and the freely estimated model. DIAMONDS dimensions depicted refer to regression weights with $p < .05$. R^2 refers to categorical model fit. Sample sizes ranged between $n_{group1} = 205$ and 234, $n_{group2} = 142$ and 170, $n_{group3} = 211$ and 230.

Abbreviations: A, Adversity; D, Duty; De, Deception; FB, Facebook-SJT; I, Intellect; M, Mating; N, Negativity; O, Positivity; PI, Personal initiative SJT; RT, Team role test; S, Sociality.

$^{*}p < .05$.

Results for description-dependent items

For four of the five description-dependent items, the relevance of perceived situation characteristics for SJT performance did not differ significantly across all groups (see Table 3). Hence, for those items, relevant perceived situation characteristics did not differ depending on whether the situation description was presented. Only for one item from the personal initiative SJT did the relevance of perceived situation characteristics for SJT performance differ across all three groups. Specifically, for Group 3, for which perceived situation characteristics were based on situation descriptions only, perceived situation characteristics did not contribute significantly to performance on this SJT item. In the two remaining groups, different DIAMONDS dimensions significantly predicted SJT performance. In fact, comparing R^2s, perceived situation characteristics made a stronger contribution to SJT responses in the condition without situation descriptions compared to the condition with situation descriptions (see Table 3). Hence, these results did not support Hypothesis 2.

Finally, we computed Spearman's rank correlations between the effect of situation descriptions on SJT performance (Cohen's *d*) and the effect of perceived situation characteristics in all three groups (R^2). This was done to test for the overall relation between the effect of description dependency and the relevance of situation construal. However, no substantial correlations were found ($r_{group1} = -.007$, $p = .983$; $r_{group2} = -.004$, $p = .991$; $r_{group1} = -.025$, $p = .940$).

4.2.3 | Ancillary analyses

Hypothesis tests revealed that situation construal serves as the underlying process behind SJT performance for all three groups. Moreover, the relevance of perceived situation characteristics for SJT performance did not differ between Groups 1 and 2 for all but one description-dependent SJT item. However, all of these items differed in difficulty between the two groups. Thus, the question arises whether situation descriptions help individuals detect a specific, correct situation construal, which in turn predicts SJT performance (i.e., correct situation construal as mediator between group membership and SJT performance).[7]

To determine the correct situation construal, we asked two subject matter experts for the work-related SJT items to rate which DIAMONDS perceptions may be helpful for identifying the right answer on these SJT items. Overall interrater reliability was intraclass correlation (ICC 2) = .71 for the work-related items (Figure S7 compares expert ratings and the mean DIAMONDS profiles in the sample across work-related SJT items).[8] We calculated profile correlations as measures of similarity between the pooled expert ratings and the test-takers' perception of situation characteristics to assess the extent of individual's correct situation construal. The mean similarity of experts and participants was $M_{group1} = .62$ ($SD_{group1} = .19$), $M_{group2} = .56$ ($SD_{group2} = .22$), and $M_{group3} = .62$ ($SD_{group3} = .22$). On average, correct perceived situation construal correlated with SJT performance with $r_{group1} = .13$ ($SD_{group1} = .10$; range: −.16 to .37), $r_{group2} = .10$ ($SD_{group2} = .09$; range: −.19 to .58), $r_{group3} = .09$ ($SD_{group3} = .08$; range: −.20 to .39). A two-way ANOVA (group × description dependency of SJT items) revealed that the profile correlations differed across groups, $F(2, 4655) = 26.33$, $p < .001$, $\eta^2 = .01$. Post hoc comparisons showed that correct situation construal was on average lower in Group 2 compared to Groups 1 and 3. Furthermore, we found a significant difference between description-dependent and description-independent SJT items, $F(1, 4655) = 57.14$, $p < .001$, $\eta^2 = .01$, in that perceived situation construal was on average more correct for description-dependent SJT items. Finally, we found a significant interaction, $F(2, 4655) = 4.72$, $p = .009$, $\eta^2 = .002$. The interaction plot (Figure S8) illustrates that the decrease in correct situation construals due to omitted situation descriptions is slightly stronger for description-dependent SJT items compared to description-independent SJT items.

We further tested whether correct situation construal mediated the relation between SJT performance and group membership. We only conducted this analysis for Groups 1 and 2 and for description-dependent SJT items, as SJT performance only differed between these groups and items. We found a significant mediating effect of correct situation construal on the relationship between the availability of situation descriptions and SJT performance for two of seven SJT items ($B_{TRT5} = -0.27$, 95% CI [−0.43, −0.14], $\beta_{TRT5} = -.09$; $B_{PI9} = -0.12$, 95% CI [−0.22, −0.02], $\beta_{PI9} = -.05$). These

effects indicate that, for those two items, omitting situation descriptions made it more difficult to correctly perceive situation construal, which mediated the decrease in SJT performance.

Finally, we aimed at gauging which specific DIAMONDS serve as predictors of SJT performance. Interestingly, for six of eight work-related SJT items, the Duty facet significantly predicted SJT performance. This was concurrent with the expert ratings. In fact, in all of these items, the situation descriptions either specified work tasks or referred to situational constraints that negatively affected overall work performance (see online Supporting Information, sample items 1 and 5). Furthermore, according to the experts, the facets Mating, Positivity, and Deception were not relevant for any of the work-related SJT items. However, hypothesis tests revealed that Positivity and Deception predicted SJT performance for three work-related SJT items.

4.3 | Discussion

Study 2 shed light on whether perceived situation characteristics can explain why some SJT items are description-dependent and some are description-independent (see Krumm et al., 2015). In line with Hypothesis 2, there were no differences in the relevance of perceived situation characteristics for SJT responses to description-independent SJT items when administered with and without situation descriptions. Thus, it may be concluded that the process underlying item responses when such SJT items are administered without situation descriptions is not different from that underlying SJT items with situation descriptions. In fact, our results suggest that both versions of the SJT items (with and without situation descriptions) involve situation construal. Notably, for three of the four description-independent items, the relevant perceived situation characteristics differed for Group 3. Hence, omitting situation descriptions did not affect the relevance of situation construal for SJT performance, but omitting response options did. Thus, our preliminary conclusion is that the relevance of situation construal for SJT performance is mostly driven by response options and not by situation descriptions.

Contrary to our theorizing, similar results were found for description-dependent items. Recall that for these items the availability versus absence of situation descriptions affected item performance (in terms of mean differences). However, the relevance of situation construal for SJT item performance was not affected by the availability or absence of situation descriptions for these items. In other words, the availability of situation descriptions may affect mean item performance (i.e., might make an SJT item easier), but add little to the actual situation dependency of the SJT item, i.e. the extent that item performance is driven by situation construal.

That being said, even though we found little difference in the relation between situation construal and SJT performance across groups, subsequent analyses suggested that participants perceived significantly less correct situation construal, as inferred from subject matter expert ratings, when situation descriptions were omitted. Hence, it was easier to correctly perceive situation construal, when situation descriptions were available. However, differences in SJT performance between groups were mediated by the groups' average correctness of situation construal for only two description-dependent SJT items. Thus, for the remaining three description-dependent SJT items, an increase in correct situation construal due to the availability of situation descriptions did not lead to improved SJT performance. This finding is in line with results by Schäpers, Mussel et al. (2019). It substantiates the interpretation that situation descriptions may be less relevant for underlying situational processes in most SJT items than previously thought.

A closer look at which specific DIAMONDS were relevant for SJT performance revealed a heterogeneous picture, with the Duty facet posing the sole exception. Duty predicted SJT performance for all SJT items addressing specific work tasks and was also rated as relevant by subject-matter experts. For all other facets, the empirical evidence and expert ratings did not coincide consistently across SJT items. Furthermore, for the knowledge SJT, subject matter experts could not agree on which perceived situation characteristics are relevant. In summary, there seemed to be no general overarching system as to which specific DIAMONDS predicted SJT performance or were rated as relevant by the experts. This is in line with Rauthmann et al. (2014) research in that situation construal seems to be to a substantial extent in the eye of the beholder.

Overall, the results of this study suggest that situation construal is an underlying driver of SJT performance, even when only response options are available. Surprisingly, this was also true for SJT items that are significantly more difficult to solve when situation descriptions are omitted (i.e., description-dependent SJT items). That is, situation construal represents the same underlying psychological process for description-dependent and description-independent SJT items. Thus, this study emphasizes the need for a conceptual differentiation between the importance of situation *descriptions* and the importance of perceived situation *characteristics* for SJT performance (i.e., omitting situation descriptions is not equivalent to omitting the situation from SJT items; see Brown et al., 2016; cf., Lievens & Motowidlo, 2016).

5 | STUDY 3

The previous two studies consistently demonstrated an empirical link between perceived situation characteristics and SJT performance. Study 3 will examine how situation construal is related to the criterion validity of SJTs.

5.1 | Methods

5.1.1 | Participants

We used G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to calculate the sample size required to ensure sufficient power ($1 - \beta = .80$) to detect a small increase in R^2 ($\Delta R^2 = .05$) in a multiple regression analysis. The a-priori power analysis revealed a necessary sample size of $N = 294$. A total of 303 participants took part in our study. Participants were recruited in 2017 and 2018 in Germany via personal contacts (e-mails), classified advertisements, online postings (job-related and private social media), and university mailing lists. As an incentive, test-takers received 10€ and were offered feedback on their results on several measures of interindividual differences. After exclusion of careless responders, $N = 285$ (174 females, 2 other; age: $M = 31.27$, $SD = 10.20$, range from 18 to 73) participants were included in the subsequent statistical analyses. On average, test-takers reported $M = 8.97$ ($SD = 9.01$) years of work experience with $M = 27.31$ ($SD = 14.34$) average weekly working hours. A total of 44% held at least an undergraduate degree, 32% had completed VET, and 24% had not completed any kind of vocational education. Most participants worked in health care (16%), academia (15%), retail (13%), or media and entertainment (10%). Additionally, we gathered 164 peer ratings for $n = 125$ participants. On average, the peer raters were $M = 34.00$ ($SD = 11.68$, range 19–76) years old and had known the participant for $M = 5.80$ ($SD = 6.45$) years. Overall, 56% of the raters were work colleagues, whereas all other raters identified as close friends or family. We also asked the peer raters to indicate on a 5-point rating scale whether they felt confident rating the participant in an occupational context ($M = 4.22$; $SD = 0.77$).

5.1.2 | Study design and materials

Study 3 was conducted as a proctored laboratory session with a median completion time of 90 minutes. Participants first completed an intelligence test and then an emotion recognition test. Afterward, similarly as in Studies 1 and 2, we administered two different SJTs as well as situation characteristic questionnaires for each SJT item. Finally, test-takers responded to several self-report measures and were asked to contact one or more work colleagues for peer-rated criterion measures.

Situational judgment tests

Similar to Study 1, we applied the SJT on personal initiative (Bledow & Frese, 2009) and the short version of the SJT measuring self-consciousness (Mussel et al., 2018). For the personal initiative SJT, we asked participants not only how they would be most likely to behave but also how they would be least likely to behave. These instructions are in line

with the test author's instructions. The reliability of this SJT was $\alpha = .65$ and $\omega = .66$. The administration of the SJT measuring self-consciousness was identical to Study 1. The reliability of this SJT was $\alpha = .69$ and $\omega = .69$.

Perceived situation characteristics

Again, the situation characteristics of all SJT items were assessed with the S8-I (Rauthmann & Sherman, 2016b), with the exception of one item for each SJT for which we applied the S8* (Rauthmann & Sherman, 2016a). In contrast to Studies 1 and 2, participants first responded to all SJT items. Afterwards, all SJT items were presented again and we then asked about perceived situation characteristics. This was done to avoid priming for the situational processing of SJT items.[9] Reliability for the eight facets of the S8* ranged from $\alpha = .50$ to .85.

Criterion measures

Several criterion measures were assessed via peer reports. We applied scales assessing peer-rated personal initiative (Frese, Fay, Hilburger, Leng, & Tag, 1997; e.g., "Actively attacks problems") on a 5-point rating scale (1 = *completely disagree*; 5 = *completely agree*) and peer-rated self-consciousness (NEO-PI-R; Ostendorf & Angleitner, 2004) on a 7-point rating scale (1 = *completely disagree*; 7 = *completely agree*). Reliability was $\alpha = .82$ for personal initiative and $\alpha = .76$ for self-consciousness. We further assessed in-role behavior (IRB; Williams & Anderson, 1991; e.g., "Performs tasks that are expected from him/her") and organizational citizenship behavior (OCBI; Williams & Anderson, 1991; e.g., "Helps others who have heavy workloads") with seven items each on a 5-point rating scale (1 = *completely disagree*; 5 = *completely agree*). We chose these broad measures of task and contextual performance to match the level of generality of the assessed perceived situation characteristics (i.e., DIAMONDS). The assessment of perceived situation characteristics in SJT items should more closely resemble real-life situational processes than SJT scores that assess specific and narrow constructs). Thus, perceived situation characteristics should also predict general measures of task and contextual performance. Reliability was $\alpha_{IRB} = .89$, $\alpha_{OCBI} = .87$. When more than one peer report was available, we calculated average ratings. ICCs for these scores ranged from .50 to .61. ICCs for the absolute rater values ranged from 0.30 to 0.67. We also assessed self-rated IRB and OCBI ($\alpha_{IRB} = .81$, $\alpha_{OCBI} = .66$).

Additional measures

In order to assess the incremental validity of perceived situation characteristics for SJT performance over and above individual differences, we also included additional predictors. First, participants completed self-report measures reflecting the SJTs' target constructs, namely personal initiative (Frese et al., 1997) and self-consciousness (Ostendorf & Angleitner, 2004). We applied the same measures that were used to assess peer-rated criteria. Reliability was $\alpha = .78$ and .70, respectively.

Second, participants completed three facets of the German version of the General Aptitude Test (Schmale & Schmidtke, 2001), which measure general mental ability. The three subtests (spatial aptitude, 40 items; numerical aptitude, 25 items; verbal aptitude, 60 items) were chosen due to their strong association with a general factor (Hunter, 1983). Reliability for the three subscales ranged from $\alpha = .82$ to .90. We computed a score for general mental ability following the test authors' instructions. Reliability of this score was $\alpha = .61$.

Third, emotional intelligence has been identified as a relevant antecedent of SJT performance (Lievens & Motowidlo, 2016). Thus, we administered the GERT-S (Schlegel & Scherer, 2016) measuring emotion recognition as an additional control variable. This test consists of 42 short video sequences in which actors express one of 14 different emotions. After each sequence, participants were asked to indicate which emotion was expressed in the video. Correct answers were scored as "1", and all other responses were scored as "0". The reliability of this test was $\alpha = .84$.

Finally, we assessed Big Five personality with the German short version of the Big Five Inventory (BFI-K; Rammstedt & John, 2005). This test measures five broad traits with a total of 21 items on a 5-point rating scale. Reliability for this test ranged from $\alpha = .67$ to .81.

5.1.3 | Data analyses

We applied path model analyses for each SJT item to simultaneously test the predictive validity on multiple criteria. Similar to Study 2, all analyses were based on single SJT items. We first tested the relation between SJT performance and the criteria and subsequently included perceived situation characteristics. We compared the two models based on R^2. We again used residual scores for the perceived situation characteristics to control for individual's general tendency to perceive multiple SJT items equally. For additional information, see Table S3.

5.2 | Results

5.2.1 | Preliminary analyses

Descriptive statistics and bivariate correlations can be found in Table 4 (see Table 1 for pooled correlations among DIAMONDS across SJT items). We again tested whether perceived situation characteristics predicted SJT performance. For 15 of 18 SJT items, we found a significant relation between DIAMONDS and SJT performance, with an average $R^2 = .05$ ($SD = .02$) for items from the personal initiative SJT and an average $R^2 = .38$[10] ($SD = .14$) for items from the self-consciousness SJT. When corrected for alpha-inflation (Bonferroni correction; $p = .05/18 \approx .0028$; Cabin & Mitchell, 2000), the link between perceived situation characteristics and SJT responses remained significant for six SJT items. In the next step, we controlled for general mental ability, emotion recognition, Big Five personality, personal initiative, and self-consciousness. Overall, this did not change the relation between perceived situation characteristics and SJT performance. On average, DIAMONDS explained $\Delta R^2 = .04$ ($SD = .01$) in personal initiative SJT performance above and beyond traditional individual difference variables. For the self-consciousness SJT, model fit increased by $\Delta R^2 = .30$ ($SD = .12$) on average. After controlling for individual differences, a significant link between perceived situation characteristics and SJT responses was found for 17 of 18 SJT items (seven items when corrected for alpha-inflation).

5.2.2 | Hypothesis tests

Overall personal initiative SJT scores predicted peer-rated personal initiative ($\beta = .193$, $p = .023$). For all other peer-rated criteria, no significant links were found. We further inspected criterion validity on the item level as perceived situation characteristics were assessed at this level. Two SJT items predicted peer-rated personal initiative and one item predicted peer-rated self-consciousness. Notably, all three of these items were from the personal initiative SJT. One item from the self-consciousness SJT predicted peer-rated OCBI.

We next added perceived situation characteristics to the analysis. For 14 of 18 SJT items, perceived situation characteristics significantly predicted at least one peer-rated criterion above and beyond SJT item performance, with average ΔR^2s of $M_{OCBI} = .080$ ($SD = .037$), $M_{IRB} = .100$ ($SD = .046$), $M_{PI} = .064$ ($SD = .030$), and $M_{SC} = .069$ ($SD = .033$). When we additionally controlled for personality, general mental ability, and emotion recognition, perceived situation characteristics exhibited similar amounts of incremental criterion validity (see Table S9). Generally, a similar picture emerged for self-rated criteria (Table S10). Thus, the results support Hypothesis 3 (for details, see Table 5).

Finally, we tested whether perceived situation characteristics mediate the relation between the personality facet measured by the SJT and SJT responses, which would be in line with person × situation interactions in situation construal models (e.g., Funder, 2016). Previous research proposed such a relation for SJTs but did not explicitly test the mediating effect (Schäpers, Mussel et al., 2019). We only found indirect effects for two items from the self-consciousness SJT, for which Positivity ($B_{N2} = 0.11$, 95% CI [0.03, 0.19], $\beta_{N2} = .10$) and Negativity ($B_{N3} = 0.07$, 95% CI [0.01, 0.14], $\beta_{N3} = .07$) mediated the relation between self-reported self-consciousness and SJT item responses.

TABLE 4 Descriptive statistics and correlations (Study 3)

	M (*SD*)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. PI SJT	5.07 (6.55)	(.65)														
2. SC SJT	1.40 (1.20)	−.13*	(.69)													
Peer-rated criteria																
3. PI peer	3.97 (0.61)	.18*	−.01	(.82)												
4. SC peer	2.86 (0.93)	−.14	.02	−.26*	(.76)											
5. OCBI peer	4.16 (0.64)	−.07	−.01	.58*	−.05	(.87)										
6. IRB peer	4.49 (0.60)	.01	.03	.58*	−.07	.76*	(.89)									
Other constructs																
7. PI self	3.65 (0.61)	.36*	−.22*	.26*	−.12	.05	.05	(.78)								
8. SC self	3.50 (0.98)	−.24*	.50*	−.05	.03*	.06	.07	−.37*	(.70)							
9. ES	3.12 (0.87)	.17*	−.26*	.06	−.28*	−.07	−.04	.32*	−.60*	(.75)						
10. E	3.56 (0.87)	.30*	−.32*	.13	−.16	.05	−.01	.38*	−.48*	.28*	(.81)					
11. C	3.75 (0.71)	.24*	−.07	.32*	−.08	.04	.14	.47*	−.19*	.19*	.21*	(.69)				
12. A	3.11 (0.84)	.08	−.08	.04	.00	.13	.10	.08	−.23*	.21*	.22*	.02	(.67)			
13. O	3.99 (0.80)	.16*	−.13*	.17	−.06	.06	.00	.30*	−.19*	.06	.25*	.05	.14*	(.77)		
14. GMA	57.75 (2.18)	−.06	−.06	−.08	−.09	.02	.00	−.09	.03	.09	−.17*	.02	−.15*	−.05	(.67)	
15. GERT-S	28.99 (4.69)	.06	.00	.00	−.18*	.08	.11	−.08	.05	−.11	.04	−.02	.01	.05	.25*	(.84)

Note. Coefficient alpha reliability is depicted on the diagonal. $n = 284–285$ for SJTs and other constructs, $n = 121–125$ for peer-rated data.
Abbreviations: A, agreeableness; C, conscientiousness; E, extraversion; ES, emotional stability; GERT-S, test of emotion recognition; GMA, general mental ability; IRB, in-role behavior; O, openness; OCBI, organizational citizenship behavior; PI, personal initiative; SC, self-consciousness.
*$p < .05$.

TABLE 5 Criterion-related validity (peer-rated) of perceived situation characteristics (Study 3)

	OCBI peer		IRB peer		PI peer		SC peer	
SJT items	DIAMONDS	ΔR^2	DIAMONDS	ΔR^2	DIAMONDS	ΔR^2	DIAMONDS	ΔR^2
SJT PI 1	S	.109	S	.084	–	.061	S	.067
SJT PI 2	–	.100	–	.043	–	.034	–	.047
SJT PI 3	–	.039	–	.065	N	.073	–	.045
SJT PI 4	–	.026	–	.039	–	.025	–	.045
SJT PI 5	D, N	.144	N	.099	D	.121		.064
SJT PI 6	I	.112	D, I, A, M	.144	–	.056	O, N	.124
SJT PI 7	–	.109	M	.123	–	.058	R̂	.084
SJT PI 8	N, S	.099	A, N, De	.181	–	.043	–	.066
SJT PI 9	–	.058	–	.038	–	.029	–	.015
SJT PI 10	D, De	.102	D, A, S	.163	D	.092	–	.049
SJT PI 11	I	.058	–	.061	–	.043	De	.136
SJT PI 12	A, De	.090	A, De	.138	De	.102	D, M	.106
SJT SC 1	–	.032	–	.064	–	.017	–	.040
SJT SC 2	O	.095	O, S	.135	O	.091	–	.057
SJT SC 3	–	.018	–	.043	–	.072	D	.078
SJT SC 4	N	.097	N	.142	N	.106	N	.105
SJT SC 5	–	.043	A	.114	–	.071	–	.022
SJT SC 6	O	.112	O, N	.124	–	.057	M	.084

Note. DIAMONDS dimensions depicted refer to regression weights with $p < .05$. ΔR^2 refers to incremental explained variance of perceived situation characteristics in criteria over and above SJT performance. $n = 125$.
Abbreviations: A, Adversity; D, Duty; De, Deception; I, Intellect; IRB, In-role behavior; M, Mating; N, Negativity; O, Positivity; OCBI, Organizational citizenship behavior; PI, Personal initiative; S, Sociality; SC, Self-consciousness.

5.3 | Discussion

The results of Study 3 demonstrated that, for almost all of the included SJT items, some facets of perceived situation characteristics predicted relevant criteria over and above SJT item responses. This is in line with Hypothesis 3 and previous results by Rockstuhl et al. (2015). Thus, SJT items have the potential to evoke relevant situation construal, which has predictive validity above and beyond SJT responses. Interestingly, situation construal had predictive validity for broad measures of contextual and task performance even when the SJT score itself was not related to these criteria. This may be interpreted as further evidence that the forced response format of SJTs may only partially capture work-relevant judgment processes, including situation construal. Directly measuring situation construal specifically captures what people think, feel, and how they make sense of a given situation. In line with substantial previous research (Rockstuhl et al., 2015), these processes turned out to be relevant for broad work-related criteria.

Additionally, Study 3 provided evidence that perceived situation characteristics capture relevant situational variance independent of individual differences. This is an important finding, as it strengthens the interpretation of perceived situation characteristics as measures of situation construal.

Contrary to situation construal theory (e.g., Funder, 2016), the relation between personality and SJT performance was not mediated by situation construal. Obtaining similar results, Schäpers, Mussel et al. (2019) concluded that situational processes may not take place in SJTs. However, our results indicate that the opposite may more likely be true: the lack of indirect effects between personality and SJT responses via perceived situation characteristics may be indicative of the complexity of situation construal and its emergence. In other words, the link between personality and situation

construal may not be linear. The notion of nonlinear interaction processes between person and situation may be fruitful for further investigations (e.g., Blum et al., 2018).

6 | GENERAL DISCUSSION

Recent studies have challenged the view of SJTs as situational measures (e.g., Krumm et al., 2015; Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010). However, most previous studies on situations in SJTs have neglected recent theorizing on person × situation interactions and more specifically, on situation construal as an underlying psychological process driving SJT performance (cf., Brown et al., 2016; Schäpers, Mussel et al., 2019). The current research therefore incorporated situation construal into a working model of SJT performance. Specifically, we tested whether situation construal affected SJT responses, whether the link between situation construal and SJT responses was contingent on the availability of situation descriptions and/or response options, and whether situation construal had incremental validity over and above SJT performance.

6.1 | Implications for theory

The first theoretical implication of this research is that situation construal is relevant for SJT performance. The three studies consistently demonstrated that situation construal predicted SJT item responses for a majority of the included SJT items. Hence, situation construal plays a pivotal role in SJT item responses. Notably, perceived situation characteristics predicted SJT responses even when controlling for individual differences (general mental ability, emotion recognition, personality, and the grand mean-centered averages of perceived situation characteristics across all SJT items). Thus, the remaining variance in perceived situation characteristics that predicted SJT responses (over and above individual differences) reflects situation-specific variance. Therefore, situation construal accounts for psychological processes underlying SJT items. According to these findings, SJTs may be understood as situational measures. This supports previous research arguing in favor of the situation dependency of SJTs (e.g., Lievens et al., 2018; Weekley et al., 2015; Westring et al., 2009) as opposed to the situation-independent perspective (e.g., Krumm et al., 2015; Lievens & Motowidlo, 2016; Schäpers, Mussel, et al., 2019).

A second theoretical implication is that the relevance of situation construal varies as a function of various, still unknown item characteristics. In all three studies, the effects of perceived situation characteristics on SJT responses differed considerably across SJT items. This finding speaks to the notion that SJT items may lie on a continuum, with some items more situational and others less situational (see Krumm et al., 2015). Interestingly, the variability in the relevance of perceived situation characteristics was not explained by whether or not a given SJT item was classified as description-dependent (due to the presence of a mean difference with vs. without situation descriptions). Hence, mean differences in SJT items that are presented with vs. without situation descriptions do not render them situational per se. Likewise, the absence of mean differences in SJT items with vs. without situation descriptions does not automatically imply that they are nonsituational. Further research is needed to identify the specific aspects of SJT items that contribute to their situation and description dependency.

Third and perhaps most remarkably, our findings suggest that response options are sufficient for situation construal to drive SJT item performance. That is, our results showed that situation construal remained relevant even when situation descriptions were omitted. In fact, our findings suggest that situation construal of SJT items may be based mostly on response options rather than on situation descriptions. This is in line with arguments that response options in SJT items also contain relevant situational cues (Fan et al., 2016; Harris et al., 2016; Melchers & Kleinmann, 2016). In fact, response options evoked the same relevant perceived situation characteristics as situation descriptions, and in some cases, response options alone were responsible for relevant situation construal. This raises the question of whether it is accurate to describe situation descriptions as low-fidelity simulations of real job situations (cf., Lievens & De Soete, 2012; Motowidlo et al., 1990; Weekley et al., 2015). However, considering situation descriptions in SJT items

superfluous may not be warranted either. Mediation analyses in Study 2 showed that—at least for some items—the availability of situation descriptions led to better situation construal and in turn to better SJT item performance. Hence, our conclusion at this point is that further research is needed to understand which types of SJT items give rise to such mediating effects and which do not. Based on the current findings, we cannot identify general rules about when and why specific perceived situation characteristics predict SJT performance.

The findings obtained in Study 3 further highlight that situation construal is pivotal for SJT validity. The finding that situation construal has incremental criterion-related validity above and beyond SJT scores is well in line with previous research (see Lievens et al., 2018; Rockstuhl et al., 2015). Hence, situation construal matters for predicting job performance. Our results further suggest that relevant situation construal for SJT responses is mainly evoked by response options. Still, in light of earlier findings by Rockstuhl et al. (2015), it seems plausible that both parts of an SJT (i.e., situation descriptions and response options) evoke distinct forms of situation construal that add to SJT validity incrementally above and beyond one another.

More generally, it should be noted that theorizing in the realm of SJTs has mostly dealt with situation descriptions—specifically with their role for SJT performance and validity (e.g., Krumm et al., 2015; Lievens & Motowidlo, 2016; Lievens & Peeters, 2008; Motowidlo et al., 1990; Weekley et al., 2015). Core theoretical principles such as behavioral consistency and correspondence between simulated content and reality essentially only referred to the situation descriptions in SJTs. In particular, Schäpers, Mussel et al. (2019) drew a direct link from the availability of situation descriptions to the relevance of situation construal. They argued that situation construal becomes less relevant as an underlying process when fewer situational cues are available (i.e., situation descriptions are omitted). Based on a manipulation of the availability of situation descriptions, the authors concluded that situation construal may have little relevance for SJTs' construct-related validity. In the current research, we explicitly tested the relation between situation construal and SJT responses and came to a more differentiated conclusion: Although situation descriptions are less relevant for SJT item responses than commonly assumed, situation construal is nevertheless a relevant underlying process of SJTs. However, for many SJT items, relevant situation construal is evoked not by situation descriptions but by the response options.

Surprisingly, response options have not been part of theories about SJT functioning. The current research suggests that response options may be a much richer source of information than previously thought. Although some previous studies have attested that response options may be informative (Kaminski, Felfe, Schäpers, & Krumm, 2019; Leeds, 2012; Leeds, 2018) and even sufficient for solving SJTs (Krumm et al., 2015; Schäpers, Lievens, Freudenstein, Hüffmeier, et al., 2019; Schäpers, Mussel, et al., 2019), they unanimously assumed that some process other than the one actually intended must be taking place. For instance, Leeds coined the term *cognitive acuity* to refer to test-takers' ability to detect subtle signs of correctness in response alternatives. The current findings suggest that response options may not only be informative for test-takers, but also stimulate the intended situation construal processes. Hence, future theorizing in the realm of SJTs might also need to account for the role of response options in SJTs.

6.2 | Implications for practice and research

6.2.1 | SJT development

Our research demonstrated an empirical link between situation construal and SJT performance, but also that SJT items lie on a continuum with respect to the relevance of situation construal. Therefore, we encourage future research to identify specific rules for when and how SJT items stimulate relevant perceived situation characteristics (e.g., when and why Duty is perceived and becomes relevant for SJT responses). In our view, such knowledge is pivotal to sufficiently capture the situational component of SJTs. Think-aloud techniques and systematic manipulations of SJT item content may be fruitful for such undertakings. Furthermore, future research is needed to examine how person variables contribute to situation construal, as our results did not support our assumption of indirect effects between personality and SJT responses.

From a more practical point of view, the current research might have an impact on SJT item development. As we found that situation construal is a driver of SJT responses, it might be valuable to think explicitly about the situation construal that should be evoked by each SJT item. Situation descriptions are usually developed using critical incident techniques (e.g., Campion et al., 2014). We suggest that practitioners include assessments of situation construal in such techniques. If subject matter experts report not only on critical situations but also how they perceive such situations, situation construal could be included from the beginning of the item development process on (see also Lievens, 2017). Subsequently, different SJT items could be clustered according to the intended constructs and to different dimensions of situation construal. Such recommendations are also in line with Trait Activation Theory (Tett & Guterman, 2000).

Closely related to the aforementioned point is research on the construct validity of SJTs. Thus far, most SJTs have struggled to meet general guidelines on convergent construct correlations (McDaniel et al., 2001) and reliability (Catano, Brochu, & Lamerson, 2012). Incorporating situation construal into the SJT development process may lead to an improvement in overall construct validity. Advanced statistical methods of variance decomposition (e.g., confirmatory factor analysis, generalizability theory, item response theory) may support this goal (see Jackson et al., 2017; Lievens et al., 2018; Westring et al., 2009).

6.2.2 | Response formats and scoring options

Another point to take into consideration is the selection of response and scoring options. Our results showed that relevant situation construal is not fully captured by test-takers' responses to SJTs. Test developers may wish to consider matching different response options with different sets of perceived situation characteristics. Furthermore, rating scales for all response options may provide more relevant information than traditional pick-the-best instructions. This may lead to a more refined measurement of situation construal, which could in turn improve SJTs' criterion validity. Alternatively, practitioners may also wish to specifically ask about test-takers' situation construal.

6.2.3 | Criterion-related validity

We call for future empirical research to enhance knowledge of why SJTs predict relevant criteria. On the one hand, this may be achieved through complementary analyses to existing meta-analytical findings that gauge the relevance of situation construal for different SJTs as a moderator of the criterion validity. On the other hand, future studies may wish to combine situation construal with other lines of research on situational effects (e.g., the frame-of-reference effect; see Shaffer & Postlethwaite, 2012) to systematically examine their effects on criterion validity.

6.2.4 | Applicant perceptions

Finally, incorporating situation construal into SJT item development could help provide more realistic job previews. If situation construal is used to create low-fidelity simulations of real-life job situations as perceived by job incumbents, responding to SJT items might help applicants more closely experience what they would experience on the job. This may further enhance HR practitioners' ability to dedicate more attention to person-job fit as a relevant criterion in the selection process. Similarly, if SJTs are used for personnel development purposes (see Thornton, Mueller-Hanson, & Rupp, 2017), additional information about test-takers' construal may help uncover the reasons for ineffective behavior.

6.3 | Limitations

First, most of the SJTs tested in this research come from a subset of SJTs with particularly distinct construct validity (e.g., personal initiative SJT, self-consciousness SJT). Thus, the generalizability of our results to all SJTs may be limited. In particular, the role of situation construal for SJT response may differ for multifaceted SJTs. However, Study 2 contained at least some items from such an SJT (TRT; Mumford et al., 2008), and the results were comparable. Moreover, our results showed that perceived situation characteristics vary across items even for unidimensional SJTs. One may

reason that if personality constructs explain moderate amounts of SJT variance, and situation construal still plays an important role, the effect may be similar or even higher for SJTs with more complex structures.

Second, we did not test the relation between perceived situation characteristics and SJT responses for video-based SJTs. Due to the higher density of situational cues in such SJTs, it may be reasonable to conclude that situation construal for video-based items is more specific and less ambiguous than for text-based items. Nonetheless, Schäpers, Lievens, Freudenstein, Hüffmeier, et al. (2019) recently demonstrated that the effect of video-based situation descriptions on SJT performance is comparable to the effect found for text-based SJTs. This may be the reason to assume that the psychological functioning of video-based SJTs is similar to that of text-based SJTs.

Third, we operationalized situation construal with the Situational Eight DIAMONDS (Rauthmann et al., 2014). This taxonomy was designed to comprehensively capture a broad range of situations (Rauthmann et al., 2014). Nevertheless, one may argue that certain facets are not suitable for situations in the work context (e.g., Mating). However, Horstmann, Rauthmann, and Sherman (2017) demonstrated large conceptual overlaps among different situation taxonomies, including taxonomies with a more work-oriented focus. The exceptions were *Typicality* (Parrigon et al., 2017) and *Lack of Stimuli* (Ziegler, 2014); hence, these may be fruitful to consider in future applications. Furthermore, these taxonomies were developed for real-life situations. In SJT items, contextual information is very restricted, which may lead to a mismatch between measures of these taxonomies and contextual information in SJT items. Nevertheless, one would expect an increase in fit between taxonomies and the presented situation descriptions to generate even larger effects than those found in our studies. Additionally, Horstmann and Ziegler (2018) recently demonstrated that the DIAMONDS exhibit substantial overlap with positive and negative affect. Thus, future research is needed to scrutinize the relation between affect and SJT responses.

Fourth, we acknowledge that, although we manipulated whether situation descriptions and response options were available as sources for situation construal in Study 2, we did not fully control for such influences on SJT performance. That is, even though situation construal in Group 3 was based solely on situation descriptions, test-takers subsequently saw all response options. An open-response format would have been the only way to prevent this. Arguably, this would have added a different type of bias in terms of the comparability of Group 3 with Groups 1 and 2. Nevertheless, we urge future research to examine the relation between situation construal and SJT performance in open-response SJTs.

Finally, we gathered peer-rated data to assess criterion-related validity in Study 3. Thus, participants may have chosen peers with a slight positive bias in their ratings. Nevertheless, situation construal predicted peer-rated criteria above and beyond SJT scores, which supports our argument that SJT scores do not capture all of the relevant situational variance. Still, we encourage future research to assess the relevance of situation construal in high-stakes settings and for supervisor ratings.

7 | CONCLUSION

This research integrated situation construal into SJT theory and thus contributed to a more fine-grained examination of SJTs as situational measures. We found that (a) situation construal significantly contributed to SJT responses, (b) situation construal was still relevant for SJT performance even when only response options were presented, and (c) situation construal explained variance in relevant criteria over and above SJT performance. Therefore, despite recent attempts to reconceptualize SJTs as context-independent measures, SJTs can still be understood as situational measures. However, situation descriptions may be less crucial for these underlying situational processes. We therefore encourage researchers and practitioners to include situation construal into item development processes and take a more theory-driven approach to constructing situation descriptions.

ACKNOWLEDGMENTS

The German Research Foundation (KR 3457/2-1) funded part of this research. We thank Mayra Borth, Mareike Breda, Alexandra Göbel, Laura Haas, Elena Harst, Christin Kalusa, Jana Kindermann, and Lilly Klimitz for their help

in collecting part of the data. We are also thankful to Cornelius König for comments on an earlier version of this manuscript. All data and R code are available on the Open Science Framework (osf.io/6kd9h).

ENDNOTES

[1] We use the terms *SJT performance* and *SJT response* interchangeably. The term *SJT scores* refers to aggregated SJT responses.

[2] Despite the overall consensus that behavior can be described using Lewin's formula (1936) as a function of both personality and situation (Fleeson & Noftle, 2008; Hogan, 2009), numerous theoretical assumptions about person × situation interactions exist (e.g., Funder, 2016; Meyer, Dalal, & Hermida, 2009; Mischel, 1968; Reis, 2008; Shoda, Mischel, & Wright, 1994; Tett & Guterman, 2000). Until quite recently, however, there was a lack of extensive theoretical descriptions of situations (Hogan, 2009; Rauthmann et al., 2014; Rauthmann, Sherman, & Funder, 2015), which was in striking contrast to the comprehensive models of personality that have long existed (e.g., Ashton & Lee, 2007; John & Srivastava, 1999). Rauthmann et al. (2014) presented such a taxonomy with situation characteristics as the key element for explaining behavior. Their work was in turn influenced by earlier conceptualizations of person × situation interactions as situation construal (e.g., Hogan, 2009; Mischel & Shoda, 1995; Reis, 2008).

[3] Because this study was the first to explicitly assess situation characteristics in SJT items, no a priori power analysis was conducted. However, we sought to obtain a total of 240 participants following general guidelines for logistic regression analysis (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996).

[4] Meta-analyses have revealed that SJTs' internal consistencies are generally low to moderate (Catano, Brochu, & Lamerson, 2012; Kasten & Freund, 2016).

[5] For all studies, we report categorical ω (Green & Yang, 2008) for SJTs.

[6] Demographics were surveyed at the end of the study. Thus, demographic data only exist for $n = 542$ participants.

[7] We thank an anonymous reviewer for suggesting this analysis.

[8] We also asked two different subject matter experts to evaluate the Facebook-SJT items. However, we only found an inter-rater reliability of $ICC2 = .27$. Thus, we only inspected ratings for the work-related SJT items.

[9] Presenting the DIAMONDS questionnaire immediately may encourage participants to inspect the situation descriptions more carefully. However, comparing the results across studies indicates that the time and placing of the DIAMONDS questionnaires had little to no effect on the relation between DIAMONDS and SJT performance. Thus, this procedure further substantiates the robustness of the effects found in Studies 1 and 2.

[10] For the SJT items measuring self-consciousness, R^2 refers to pseudo-R^2 in lavaan (Rosseel, 2012) due to the categorical nature of the dependent variable.

ORCID

Jan-Philipp Freudenstein https://orcid.org/0000-0002-9029-5003
Philipp Schäpers https://orcid.org/0000-0002-8270-5105

REFERENCES

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*, 150–166. https://doi.org/10.1177/1088868306294907

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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

Blum, G. S., Rauthmann, J. F., Göllner, R., Lischetzke, T., Schmitt, M., & Kandler, C. (2018). The nonlinear interaction of person and situation (NIPS) model: Theory and empirical evidence. *European Journal of Personality*, *32*, 286–305. https://doi.org/10.1002/per.2138

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-3456-0

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: Perspectives on Science and Practice*, *9*, 38–42. https://doi.org/10.1017/iop.2015.113

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

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. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). London, UK: Routledge. https://doi.org/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. https://doi.org/10.1080/08959285.2014.929693

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. https://doi.org/10.1111/j.1468-2389.2012.00604.x

Chen, L., Fan, J., Zheng, L., & Hack, E. (2016). Clearly defined constructs and specific situations are the currency of SJTs. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *9*, 34–38. https://doi.org/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. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Crook, A. E. (2016). Unintended consequences: Narrowing SJT usage and losing credibility with applicants. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *9*, 59–63. https://doi.org/10.1017/iop.2015.118

Debusscher, J., Hofmans, J., & De Fruyt, F. (2016). Do personality states predict momentary task performance? The moderating role of personality variability. *Journal of Occupational and Organizational Psychology*, *89*, 330–351. https://doi.org/10.1111/joop.12126

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138. https://doi.org/10.1037/1082-989X.12.2.121

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: Perspectives on Science and Practice*, *9*, 43–47. https://doi.org/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. https://doi.org/10.3758/bf03193146

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, *75*, 825–862. https://doi.org/10.1111/j.1467-6494.2007.00458.x

Fleeson, W., & Noftle, E. (2008). The end of the person-situation debate: An emerging synthesis in the answer to the consistency question. *Social and Personality Psychology Compass*, *2*, 1667–1684. https://doi.org/10.1111/j.1751-9004.2008.00122.x

Frese, M., Fay, D., Hilburger, T., Leng, K., & Tag, A. (1997). The concept of personal initiative: Operationalization, reliability and validity in two German samples. *Journal of Occupational and Organizational Psychology*, *70*, 139–161. https://doi.org/10.1111/j.2044-8325.1997.tb00639.x

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. https://doi.org/10.1177/0963721416635552

Green, S. B., & Yang, Y. (2008). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155–167. https://doi.org/10.1007/s11336-008-9099-3

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: Perspectives on Science and Practice*, *9*, 23–28. https://doi.org/10.1017/iop.2015.110

Harvey, R. J. (2016). Scoring SJTs for traits and situational effectiveness. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *9*, 63–71. https://doi.org/10.1017/iop.2015.119

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64–78. https://doi.org/10.1037/1082-989x.2.1.64

Hemmert, G. A. J., Schons, L. M., Wieseke, J., & Schimmelpfennig, H. (2016). Log-likelihood-based pseudo-R^2 in logistic regression. *Sociological Methods & Research*, *47*, 507–531. https://doi.org/10.1177/0049124116638107

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

Horowitz, J. L. (1982). Evaluation of usefulness of two standard goodness-of-fit indicators for comparing non-nested random utility models. *Transportation Research Record*, *874*, 19–25.

Horstmann, K. T., Rauthmann, J. F., & Sherman, R. A. (2017). Measurement of situational influences. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences* (pp. 465–484). London, UK: SAGE. https://doi.org/10.4135/9781526451163.n21

Horstmann, K. T., & Ziegler, M. (2018). Situational perception and affect: Barking up the wrong tree? *Personality and Individual Differences*. https://doi.org/10.1016/j.paid.2018.01.020

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Hove, UK: Routledge. https://doi.org/10.4324/9780203852279

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 service*. Retrieved from https://eric.ed.gov/?id=ED236166

Jackson, D. J. R., 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. https://doi.org/10.1111/joop.12151

Jansen, A., Melchers, K. G., Lievens, F., Kleinmann, M., Brändli, M., Fraefel, L., & König, C. J. (2013). Situation assessment as an ignored factor in the behavioral consistency paradigm underlying the validity of personnel selection procedures. *Journal of Applied Psychology*, *98*, 326–341. https://doi.org/10.1037/a0031257

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York, NY: Guilford.

Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology*, *95*, 54–78. https://doi.org/0.1037/a0017286

Kaminski, K., Felfe, J., Schäpers, P., & Krumm, S. (2019). A closer look at response options: Is judgment in situational judgment tests a function of the desirability of response options? *International Journal of Selection and Assessment*. https://doi.org/10.1111/ijsa.12233

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. https://doi.org/10.1027/1015-5759/a000250

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*, 399–417. https://doi.org/10.1037/a0037674

Leeds, J. P. (2012). The theory of cognitive acuity: Extending psychophysics to the measurement of situational judgment. *Journal of Neuroscience, Psychology, and Economics*, *5*, 166–181. https://doi.org/10.1037/a0027294

Leeds, J. P. (2018). Applying cognitive acuity theory to the development and scoring of situational judgment tests. *Behavior Research Methods*, *50*, 2215–2225. https://doi.org/10.3758/s13428-017-0988-1

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*, 411–431. https://doi.org/10.1080/00273170802285743

Lewin, K. (1936). *Principles of topological psychology*. New York, NY: McGraw-Hill.

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing*, *17*, 269–276. https://doi.org/10.1080/15305058.2017.1309857

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology*, *93*, 268–279. https://doi.org/10.1037/0021-9010.93.2.268

Lievens, F., & De Soete, B. (2012). Simulations. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 383–410). Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199732579.013.0017

Lievens, F., Lang, J., De Fruyt, F., Corstjens, J., Van de Vijver, M., & Bledow, R. (2018). The predictive power of people's intraindividual variability across situations: Implementing whole trait theory in assessment. *Journal of Applied Psychology*, *103*, 753–771. https://doi.org/10.1037/apl0000280

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

Lievens, F., & Peeters, H. (2008). Impact of elaboration on responding to situational judgment test items. *International Journal of Selection and Assessment*, *16*, 345–355. https://doi.org/10.1111/j.1468-2389.2008.00440.x

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

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. https://doi.org/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. https://doi.org/10.1017/iop.2015.115

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730–740. https://doi.org/10.1037//0021-9010.86.4.730

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, *9*, 103–113. https://doi.org/10.1111/1468-2389.00167

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455. https://doi.org/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: Perspectives on Science and Practice*, *9*, 29–34. https://doi.org/10.1017/iop.2015.111

Meyer, R. D., Dalal, R. S., & Hermida, R. (2009). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*, 121–140. https://doi.org/10.1177/0149206309349309

Mischel, W. (1968). *Personality and assessment*. New York, NY: Psychology Press.

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. Retrieved from https://doi.org/1995-25136-001

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. https://doi.org/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. https://doi.org/10.1037/0021-9010.75.6.640

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*, 749–761. https://doi.org/10.1037/0021-9010.91.4.749

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 57–81). Mahwah, NJ: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203774878

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. https://doi.org/10.1037/0021-9010.93.2.250

Mussel, P., Gatzka, T., & Hewig, J. (2018). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment*, *34*, 328–335. https://doi.org/10.1027/1015-5759/a000346

Mussel, P., Schäpers, P., Schulz, J.-P., Schulze, J., & Krumm, S. (2017). Assessing personality traits in specific situations: What situational judgment tests can and cannot do. *European Journal of Personality*, *31*, 475–476. https://doi.org/10.1002/per.2119

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620. https://doi.org/10.1207/s15328007sem0904_8

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68. https://doi.org/10.1016/j.jrp.2015.09.001

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

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (NEO-PI-R) [NEO Personality Inventory based on Costa and McCrae, revised version (NEO-PI-R)]*. Göttingen, Germany: Hogrefe.

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. https://doi.org/10.1037/pspp0000111

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*, 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

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

Ployhart, R. E., & MacKenzie, W. I., Jr. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA handbooks in psychology. APA handbook of industrial and organizational psychology, Vol. 2. Selecting and developing members for the organization* (pp. 237–252). Washington, DC: American Psychological Association. https://doi.org/10.1037/12170-008

Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2018). Validation of the short and extra-short forms of the Big Five Inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000481

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K). [Short version of the Big Five Inventory]. *Diagnostica*, *51*, 195–206. https://doi.org/10.1026/0012-1924.51.4.195

Rauthmann, J. F. (2015). Structuring situational information. A road map of the multiple pathways to different situational taxonomies. *European Psychologist*, *20*, 176–189. https://doi.org/10.1027/1016-9040/a000225

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. https://doi.org/10.1037/a0037250

Rauthmann, J. F., & Sherman, R. A. (2016a). Measuring the Situational Eight DIAMONDS characteristics of situations: An optimization of the RSQ-8 to the S8*. *European Journal of Psychological Assessment*, *32*, 155–164. https://doi.org/10.1027/1015-5759/a000246

Rauthmann, J. F., & Sherman, R. A. (2016b). Ultra-brief measures for the Situational Eight DIAMONDS domains. *European Journal of Psychological Assessment*, *32*, 165–174. https://doi.org/10.1027/1015-5759/a000245

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations. *European Journal of Personality*, *29*, 363–381. https://doi.org/10.1002/per.1994

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. https://doi.org/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. https://doi.org/10.1177/1088868308321721

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. https://doi.org/10.1037/a0038098

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

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis* (pp. 233–247). Boston, MA: Springer. https://doi.org/10.1007/978-1-4615-4603-0_17

Saucier, G., Bel-Bahar, T., & Fernandez, C. (2007). What modifies the expression of personality tendencies? Defining basic domains of situation variables. *Journal of Personality*, *75*, 479–503. https://doi.org/10.1111/j.1467-6494.2007.00446.x

Schäpers, P., Lievens, F., Freudenstein, J.-P., Hüffmeier, J., König, C. J., & Krumm, S. (2019). Removing situation descriptions from situational judgment test items: Does the impact differ for video-based versus text-based formats? *Journal of Occupational and Organizational Psychology*. https://doi.org/10.1111/joop.12297

Schäpers, P., Lievens, F., Freudenstein, J.-P., Schulze, J., König, C. J., & Krumm, S. (2019, May). *Which kind of situational information is needed to make situational judgment tests situational*? Paper presented at the 19th European Association of Work and Organizational Psychology (EAWOP) Congress, Turin, Italy.

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 reactions. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000457

Schlegel, K., & Scherer, K. R. (2016). Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation. *Behavior research methods*, *48*, 1383–1392. https://doi.org/10.3758/s13428-015-0646-4

Schmale, H., & Schmidtke, H. (2001). *Berufseignungstest (BET) [General Aptitude Test Battery (GATB)]*. Bern, Switzerland: Huber.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology*, *80*, 607–620. https://doi.org/10.1037/0021-9010.80.5.607

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, *65*, 445–494. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922–932. https://doi.org/10.1037/0003-066x.44.6.922

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. https://doi.org/10.1037/pspp0000036

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, *67*, 674–687. https://doi.org/10.1037/0022-3514.67.4.674

Snijders, T. A. B. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality & Quantity*, *30*, 405–426. https://doi.org/10.1007/BF00170145

St-Sauveur, C., Girouard, S., & Goyette, V. (2014). Use of situational judgment tests in personnel selection: Are the different methods for scoring the response options equivalent? *International Journal of Selection and Assessment*, *22*, 225–239. https://doi.org/10.1111/ijsa.12072

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*, 500–517. https://doi.org/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. https://doi.org/10.1006/jrpe.2000.2292

Thornton, III, G. C., Mueller-Hanson, R. A., & Rupp, D. E. (2017). *Developing organizational simulations. A guide for practitioners, students, and researchers*. New York, NY: Routledge. https://doi.org/10.4324/9781315652382

Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis*, *22*, 537–557. https://doi.org/10.1016/0167-9473(96)00004-7

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 295–322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

Weekley, J. A., & Ployhart, R. E. (2006a). An introduction to situational judgment testing. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgment Tests. Theory, measurement and application* (pp. 1–11). Mahwah, NJ: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203774878

Weekley, J. A., & Ployhart, R. E. (2006b). Situational judgment: Some suggestions for future science and practice. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgment Tests. Theory, measurement and application* (pp. 345–351). Mahwah, NJ: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203774878

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157–182). Mahwah, NJ: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203774878

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*, 372–376. https://doi.org/10.1037/h0026244

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. https://doi.org/10.1080/08959280802540999

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*, 291–309. https://doi.org/10.1080/08959280802137820

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. https://doi.org/10.1177/014920639101700305

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48. https://doi.org/10.1016/j.jml.2013.12.003

Zaccaro, S. J., Green, J. P., Dubrow, S., & Kolze, M. (2018). Leader individual differences, situational parameters, and leadership outcomes: A comprehensive review and integration. *Leadership Quarterly*, *29*, 2–43. https://doi.org/10.1016/j.leaqua.2017.10.003

Ziegler, M. (2014). *B5PS. Big Five inventory of personality in occupational situations*. Mödling, Austria: Schuhfried.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Freudenstein J-P, Schäpers P, Roemer L, Mussel P, Krumm S. Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. 2020;1–32. https://doi.org/10.1111/peps.12385