Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

Underlying Psychological Processes of Situational Judgment Tests. Towards a Theory-Driven Integration of Person-Situation Interactions

Dissertation

zur Erlangung des akademischen Grades Doktor der Philosophie (Dr. phil.)

vorgelegt von

M.Sc.
Jan-Philipp Freudenstein

Berlin, Juni 2020

Erstgutachter
Prof. Dr. Stefan Krumm

Zweitgutachter
Prof. Dr. Patrick Mussel

Tag der Disputation: 22.07.2020

Gemeinsame Promotionsordnung zum Dr. phil. / Ph.D. der Freien Universität Berlin vom 2. Dezember 2008 (FU-Mitteilungen 60/2008)

Table of Contents

Motivation and Acknowledgements

Studying psychology, I quickly became fascinated by the concept of psychology as an empirical science. Psychologists are particularly interested in latent constructs such as intelligence, personality traits, or attitudes and their relation to behavior or other criteria. None of these constructs can be directly observed. So, finding indicators that assess these constructs is an immensely challenging yet exciting task. To date, I am convinced that psychological assessment is one of the most important premises for high-quality psychological research.

During my undergraduate studies in business psychology, I was mainly interested in personnel selection. Later, I gained deeper insights into various fields of psychology, which sparked my interest for broader problems of psychological assessments and applied psychometrics. Researching Situational Judgment Tests appealed to both these interests. These tests were primarily designed for personnel selection but comprised interesting features, namely the situational component, that represented potential assessment opportunities for applications beyond selection purposes. Just before I started my PhD program, the construct-related validity of SJTs was described as "hot mess" (McDaniel et al., 2016, p. 47) and Stefan Krumm uncovered that Situational Judgment Tests may not work as they were intended to (Krumm et al., 2015). This pile of sharps spurred my ambition to make sense of this method and to contribute to sound psychological assessments.

Beyond the opportunity for an interesting research project, this PhD program provided an inconceivable learning experience. In the last three and a half years, not a single day went by that I did not learn something new. I learned a lot about Situational Judgment Tests, new methods, interesting fields of research, Open Science, and academic writing. I read captivating articles, had heated debates, and cursed the bureaucratic machinery. But this listing barely scratches the surface! I owe this experience to my supervisors, colleagues and friends, contributors, and people I met during this time. But by far the biggest element to the success of this time was the tremendous freedom in research I enjoyed. I am thankful to all who supported me during this time and helped me to achieve this goal:

Maren, Ida, Stefan, Patrick, Philipp, Julian, Nathalie, Jantje, Nomi, Nico, Alex, Mareike, Sigrun, Sibylle, Katharina, Talea, Johannes, Julia, Jennifer, Lena, Melanie, Selina, Marius, Jan, Nita, Alexandra, Theresa, Martin, Oliver, Klaus, Heike, Joachim

Abstract

In recent years, more and more psychological assessments aimed at capturing interactions between the person and situations. Situational Judgment Tests (SJTs) are built on a similar premise, as they were designed as low-fidelity simulations of situations. These tests incorporate short situation descriptions with several behavioral response options. However, the validity and underlying psychological processes of SJTs generally remained subject to debate as a growing body of research suggested that SJTs may reflect context-independent measures. Within this debate, other scholars argued in favor of the relevance of person-situation processes for SJT responses. So far, sufficient evidence that unravels the true underlying processes of SJTs is missing. This dissertation aims at closing this gap and at contributing to a deeper understanding of SJTs as psychological assessment methods. Four empirical research papers provide theory-driven insights on context-independent and person-situation processes as potential determinants of SJT responses. First, the construct-related validity of Implicit Trait Policies is examined and therefore the notion of SJTs as context-independent measures. Next, situation construal (i.e., the perception of situations), and processes postulated by Trait Activation Theory are considered as relevant theoretical underpinnings for SJTs. Results overall supported the relevance of person-situation interactions as underlying processes and particularly challenged SJTs as measures of Implicit Trait Policies. Especially situation construal explained SJT responses consistently across three studies. However, the results also showed that not situation descriptions but response options were often crucial for relevant person-situation processes as captured in SJT responses. This lack of impact of situation descriptions also potentially limited the explanatory power of Trait Activation Theory in the context of SJT items. The results are discussed in regard to the debate about underlying processes of SJT responses. All in all, these studies raise the question whether key design features of common SJTs (i.e. situation descriptions and response options) are optimally developed for the assessment of person-situation interactions. The final paper of this dissertation introduces Standardized State Assessment as narrower and theory-driven methodological framework for the assessment of psychological states in hypothetical situations. Limitations of this dissertation, as well as implications for research and practice of psychological assessments based on situation descriptions are discussed.

Zusammenfassung

Die Berücksichtigung psychologischer Prozesse, die die Interaktion zwischen Personeneigenschaften und Situationen widerspiegeln, hat in den letzten Jahren für die psychologische Diagnostik an Bedeutung gewonnen. Dieser Prozess zeigt sich auch in Situational Judgment Tests (SJTs), die ursprünglich als simulationsbasiertes Verfahren entwickelt wurden. Diese Tests enthalten kurze Situationsbeschreibungen und mehrere verhaltensbasierte Antwortoptionen. Die Validität und die zugrundeliegenden psychologischen Prozesse von SJTs sind bislang allerdings nicht abschließend geklärt. Insbesondere neure Studien legen nahe, dass SJTs kontextunabhängige Messungen repräsentieren. Gleichzeitig existieren mehrere Argumente, die für situationsabhängige Prozesse in SJTs sprechen. Bislang fehlen jedoch ausführliche und abschließende Untersuchungen dieser Prozesse. Diese Dissertation möchte diese Lücke schließen und zu einem tieferen Verständnis von SJTs als Methode der psychologischen Diagnostik beitragen. Anhand von vier empirischen Artikeln werden theoriegeleitete Annahmen über kontext- und situationsabhängige Prozesse, die SJTs zugrunde liegen könnten, untersucht. Zunächst steht die Konstruktvalidität von Implicit Trait Policies im Vordergrund, die als erklärendes Konstrukt für SJTs als kontextunabhängige Messungen vorgebracht wurden. Weiterhin werden die Situationswahrnehmung und zentrale Aspekte der Trait Activation Theory als relevantes theoretisches Gerüst für SJTs untersucht. Die Ergebnisse unterstützen insgesamt die Relevanz situationsabhängiger Prozesse für SJTs und Zweifeln insbesondere an der Validität von Implicit Trait Policies. Vor allem die Situationswahrnehmung von SJT Items konnte das Antwortverhalten konsistent über drei Studien hinweg vorhersagen. Allerdings zeigte sich auch, dass hauptsächlich Antwortoptionen und nicht Situationsbeschreibungen entscheidend für situationsbasierte Prozesse in SJTs sind. Dies könnte auch die fehlende Relevanz der Trait Activation Theory für SJTs erklären. Die Ergebnisse werden im Kontext der Debatte über zugrundeliegende Prozesse von SJTs betrachtet. Insgesamt werfen die Ergebnisse die Frage auf, ob bisherige Konstruktionsweisen von SJTs (d.h. Situationsbeschreibungen und Antwortoptionen) eine optimale Erfassung von Interaktionen zwischen Personeneigenschaften und Situationen ermöglicht. Der letzte Artikel dieser Dissertation schlägt Standardized State Assessment als enger gefasstes und theoriegeleitetes, methodisches Modell für die Messung psychologischer Momentanzustände vor. Einschränkungen dieser Dissertation, sowie auch Konsequenzen für die Anwendung von und Forschung über psychologische Diagnostik mittels Situationsbeschreibungen werden diskutiert.

Chapter 1

Introduction

Psychological science strives to understand and explain individual behavior. In personality psychology, different perspectives about underlying processes of behavior led to the person-situation debate, in which scholars argued in favor of either stable personality traits or situational influences as underlying determinants of behavior (e.g., Epstein, 1979; Epstein & O'Brien, 1985; Mischel, 1968). For example, the five-factor theory of personality postulates five stable traits that structure human personality and serve as the basis for peoples´ actions (e.g., Digman, 1990; McCrae & Costa, 1987). Other theories argue that behavior is predominantly influenced by situation characteristics since behavior is rather unstable over time and personality traits only moderately predict actual behavior (e.g., Mischel, 1968). Today, it is widely accepted that both person characteristics and situation characteristics influence individual behaviors (e.g., Baumert et al., 2017; Fleeson & Noftle, 2008; Funder, 2016; Mischel, 1979; Mischel & Shoda, 1995; Steyer et al., 1992). That is, individual behavior is, to a certain extent, consistent across situations or measurement occasions while situation-specific components of behavior also exist (e.g., Fleeson, 2001; Steyer & Schmitt, 1990). In fact, almost a century ago Kantor (1924) already outlined human behavior as an individuals' interaction with occurring situations. In a similar vein, Lewin's (1936) infamous function of behavior incorporated the person and the situation.

Building on this proposition of person-situation processes, psychological assessment progressively considered influences of both person and situation characteristics. For example, the use of ambulatory assessment increased considerably in the past years (Hamaker & Wichers, 2017). Ambulatory assessment is an umbrella term for methods that examine psychological constructs within an individual's real environment across several occasions or situations via daily diary reports or repeated measurements throughout the day (e.g., Hofmans et al., 2019; Trull & Ebner-Priemer, 2014). Thus, they enable researchers to disentangle influences of person characteristics, situation characteristics and their interactions on the expression of the examined constructs. This assessment approach has not only been applied to personality research (e.g., Bleidorn, 2009; Fleeson, 2001; Rauthmann et al., 2016; Wilson et al., 2017) but also to clinical assessment (e.g., A. J. Fisher & Boswell, 2016; Zimmermann et al., 2019) or work and organizational psychology (e.g., C. D. Fisher & To, 2012; Ohly et al., 2010; Sonnentag & Binnewies, 2013).

Similar to the methodology of assessment in real-life and naturally occurring situations, simulation-based methods have been developed for personnel selection (Lievens & De Soete, 2012; Sackett & Lievens, 2008; Thornton III & Rupp, 2004; Weekley et al., 2015). These methods, such as assessment center exercises (e.g., role play or group discussion), try to simulate behavior in work-related tasks or work-related situations in order to predict future job performance (Lievens & De Soete, 2012; Sackett & Lievens, 2008). Although person-situation processes were not always explicitly incorporated in

simulation methods, an increasing body of research integrated these assumptions to understand underlying processes of personnel selection methods and to enhance their development (e.g., Haaland & Christiansen, 2002; Jansen et al., 2013; Oliver et al., 2016).

Situational Judgment Tests (SJTs) are similar methods that stem from the field of personnel selection (Motowidlo et al., 1990) and are a particular focus of this dissertation. In contrast to the above-mentioned methods, no real behavior in specific situations is observed. Rather, SJTs consist of short descriptions of hypothetical situations and provide several behavioral response options (see Figure 1 for an example item; Corstjens et al., 2017; Lievens et al., 2020; McDaniel & Nguyen, 2001). Test-takers are asked to pick the response option that resembles how they should or would behave in the given situation (McDaniel & Nguyen, 2001). As these tests do not measure real-life behavior, SJTs are also described as low-fidelity simulations (Motowidlo et al., 1990; Weekley et al., 2015). Nevertheless, situation descriptions in SJTs are typically defined as the essential test element (e.g., Motowidlo et al., 1990; Weekley et al., 2006). Accordingly, some scholars argued that psychological processes underlying SJT performance may be equivalent to those in real-life situations (e.g., Brown et al., 2016; Harris et al., 2016). On the contrary, other scholars showed that situation descriptions in SJTs are often less relevant to the response behavior than previously assumed (e.g., Krumm et al., 2015; Lievens & Motowidlo, 2016; Schäpers et al., 2019). Therefore, they argued that rather context-independent processes may underly SJT performance.

Figure 1

Sample SJT Item

You are under enormous pressure to accomplish your task on time. Yesterday, new trainees started in your department. They are unfamiliar with the workflow in your department. You have to interrupt your work to answer trainees' questions and to correct their mistakes. You are expected to do both, to finish your work on time and take care of trainees.

What would you do?

A I tell the trainees that I am available after work to answer their questions.

B I openly say that I cannot take care of the trainees and work for better initial training of the trainees.

C I send the trainees to my colleagues when they have questions.

D I try to get by without becoming stressed and worn out.

Notes. Item taken from the Personal Initiative SJT (Bledow & Frese, 2009, p. 223).

Table 1

Overview of Studies Included in this Dissertation.

Implicit Trait Policies	
Appendix A	**Freudenstein, J.-P.**, & Krumm, S. (2020). *Developing a short-form situational judgment test to assess implicit trait policies for agreeableness*. OSF Preprints. https://doi.org/10.31219/osf.io/kax7n
Appendix B	**Freudenstein, J.-P.**, Remmert, N., Reznik, N., & Krumm, S. (2020). *English translation of the teamwork situational judgment test (SJT-TW)* [Manuscript submitted for publication].
Chapter 2	**Freudenstein, J.-P.**, Mussel, P., & Krumm, S. (2020). *On the construct-related validity of implicit trait policies* [Manuscript prepared for publication].
Person-Situation Processes in SJTs	
Appendix C	Schäpers, P.★, **Freudenstein, J.-P.★**, Mussel, P., Lievens, F., & Krumm, S. (2020). Effects of Situation Descriptions on the Construct-Related Validity of Construct-Driven Situational Judgment Tests. *Journal of Research in Personality*. https://doi.org/10.1016/j.jrp.2020.103963
Chapter 3	**Freudenstein, J.-P.**, Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. https://doi.org/10.1111/peps.12385
Chapter 4	**Freudenstein, J.-P.**, Schäpers, P., & Krumm, S. (2020). *The influence of situational strength on the relation of personality and SJT performance* [Manuscript prepared for publication].
Standardized State Assessment	
Chapter 5	**Freudenstein, J.-P.**, Schulze, J., Schäpers, P., Mussel, P., & Krumm, S. (2020). *Standardized state assessment: A methodological framework to assess person-situation processes in hypothetical situations* [Manuscript prepared for publication].

Notes. ★ shared first authorship.

So far, sufficient evidence that unravels the true underlying processes of SJTs is missing. This dissertation aims at closing this gap and at contributing to a deeper understanding of SJTs as psychological assessment method. Such insights are pivotal for insights on how SJTs function as predictors of behavior and relevant criteria as well as for the development of SJTs for use in personnel selection. A particular focus of this dissertation is the assessment of person-situation processes. Recently, Lievens (2017a) argued that SJTs may be adequate tools to examine these processes, as they enable between-subject comparisons in specific situations. However, whether or not SJTs hold up to this premise is subject of debate So, SJTs have to be viewed in context of theoretical advances on person-situation processes and general approaches to asses these processes. Accordingly, this Chapter briefly outlines these literatures before taking a closer look at different perspectives on underlying processes of SJTs. Based on this review, I propose a working model of SJTs that provides a falsifiable structure of psychological processes underlying SJTs. In brief, this working model combines context-independent processes (e.g., Lievens & Motowidlo, 2016) and person-situation processes (e.g., Brown et al., 2016) as underlying functions of SJTs. The empirical studies reported in this dissertation examine key assumptions of the working model (see Table 1). Finally, I build on these results to propose a methodological framework for the assessment of person-situation processes with hypothetical situations that are compliant with theoretical foundations of real-life behavior. Overall, this dissertation contributes to a more fine-grained knowledge about how SJTs, and situation descriptions of hypothetical situations in general, may serve as valid assessment tool for research on person-situation processes and selection purposes in practice.

Person-Situation Processes

Most contemporary theories on individual behaviors incorporate person-situation processes (e.g., Fleeson & Jayawickreme, 2015; Mischel & Shoda, 1995; Steyer et al., 1999; Tett & Guterman, 2000). Especially Whole Trait Theory (Fleeson & Jayawickreme, 2015) may be a useful framework to integrate several theoretical propositions about individual behavior. According to Whole Trait Theory personality traits may be separated into a descriptive part and an explanatory part (Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019). The descriptive part of traits is defined as the density distribution of trait-relevant states in the form of momentary thoughts, feelings, and behaviors (Fleeson, 2001; Fleeson & Jayawickreme, 2015). For example, individuals may vary in the degree to which they act sociable and outgoing on different occasions. Accordingly, each instance of these expressions may be understood as state, which taken together represent intraindividual distributions of trait-relevant states (Fleeson, 2001;

Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019). The mean of these distributions is thought to reflect a stable, general tendency of trait-relevant thoughts, feelings and behaviors (Fleeson, 2001). Thus, the mean of state distributions reflects a dispositional tendency as described in trait theories (e.g., Ashton & Lee, 2007; Digman, 1990). Importantly, other distribution parameters such as the intraindividual variability around the mean comprises stable information about individuals as well (Fleeson, 2001; Fleeson & Jayawickreme, 2015; Jones et al., 2017). For example, several studies demonstrated temporal stability of the intraindividual variability of states (Fleeson, 2001; Jones et al., 2017). Thus, deviations from an individual's general tendency in trait expressions are psychologically meaningful (see also Steyer et al., 1992; Steyer & Schmitt, 1990).

The explanatory part of traits in Whole Trait Theory reveals causal mechanisms of individual behaviors (Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019). It reflects the notion that personality research "should identify the intraindividual psychological processes that explain variation of behaviour across situations as well as the systematic inter-individual differences in those processes that explain variation in behavior across individuals" (Baumert et al., 2017, p. 515). Beyond person characteristics (i.e., traits), situation characteristics and the interaction of both person and situation characteristics have been considered as such explanatory links (e.g., Dweck, 2017; Funder, 2016; Meyer et al., 2010; Mischel & Shoda, 1995; Tett & Guterman, 2000; Funder, 2006). For example, Trait Activation Theory posits that a specific situation must generally be trait-relevant so that trait relevant behaviors may be observed (Tett & Burnett, 2003; Tett & Guterman, 2000). That is, the situation must provide the opportunity to express different extents of trait-relevant behaviors. The opportunity to be talkative, for example, is restricted when attending a lecture. Tett and Guterman (2000) demonstrated that individuals behave more consistent with their general trait tendencies in trait-activating situations. Beyond trait activation, situational strength further influences the relation of trait dispositions and behavior (Meyer et al., 2010; Mischel, 1977; Tett & Guterman, 2000). Situational strength is defined as situational attributes that influence the "desirability of potential behaviors" (Meyer et al., 2010, p. 122). In stronger situations, appropriate behaviors are heavily determined by the situation and less driven by personality dispositions (Meyer et al., 2010; Tett & Guterman, 2000).

Importantly, several scholars emphasized the psychological relevance of situation perceptions (Funder, 2016; Meyer et al., 2014; Mischel & Shoda, 1995; Rauthmann et al., 2014; Rauthmann, Sherman, & Funder, 2015; Reis, 2008). This perception of a situation – the situation construal – affects individuals' behaviors. Funder (2016) argued that direct links between person characteristics (e.g., personality traits, abilities) as well as links between situation attributes and behavior exist. The link between situation attributes and behavior stems from objective entities such as rules or incentives, whereas situations obtain psychological relevance as a result of an individual construal. Hence,

situation construal reflects the psychological representation of person-situation processes and thus explaining interindividual and intraindividual differences in behavior (Funder, 2016; Rauthmann, Sherman, Nave, et al., 2015).

The increased attention on person-situation processes also sparked research on the conceptualization of situations (Rauthmann, Sherman, & Funder, 2015). Generally, situational information can be described by three different concepts (Rauthmann, 2015; Rauthmann, Sherman, & Funder, 2015). First, situation cues objectively describe situations (Rauthmann, 2015; Saucier et al., 2007). As such, cues comprise answers to five questions: "Who is with you? Which objects are around you? What is happening? Where are you? When is this happening?" (Rauthmann, Sherman, & Funder, 2015, p. 364). Similarly, Saucier et al. (2007) found that cues that describe locations, associations, and activities are especially useful to describe personality-relevant situations. Second, situation characteristics are individual perceptions of situation cues (Rauthmann et al., 2014; Rauthmann, Sherman, & Funder, 2015). Thus, they "capture the psychologically important meanings" of situations – the situation construal – which drives behavior and state expressions (Rauthmann, Sherman, & Funder, 2015, p. 364). Several taxonomies of relevant situation characteristics have been proposed (see Horstmann et al., 2017 for an overview). Importantly, situation characteristics have been demonstrated to predict behavior above and beyond personality traits (Parrigon et al., 2016; Rauthmann et al., 2014; Sherman et al., 2015). Finally, situation classes summarize several situations based on either cues or characteristics (Rauthmann, Sherman, & Funder, 2015).

Psychological Assessment of Person-Situation Processes

To comply with recent theories, person-situation processes haven been incorporated into psychological assessment. Especially ambulatory assessment provides the advantage of measuring states, behaviors or other constructs of interest repeatedly in a large number of real-life situations and different environments (e.g., C. D. Fisher & To, 2012; Trull & Ebner-Priemer, 2014). Due to methodological developments, adequate analyses of person-situation processes in these assessments became possible. For example, multilevel regression models allow to differentiate between variance in individual responses due to specific situations or measurement occasions and individuals (e.g., Hox, 2010; Nestler et al., 2019). Similarly, Latent State-Trait Theory uses latent variable models to distinguish between consistent variance components across occasions (i.e., the trait) from occasion-specific variance components (i.e., the state; Steyer et al., 1992, 1999). Latent state-trait models additionally separate measurement error from reliable measurement variance (Geiser et al., 2017; Steyer et al., 1999). Despite these methodological advances, validity and psychometric properties are typically not examined in ambulatory assessments (C. D. Fisher & To, 2012; Hofmans et al., 2019; Horstmann & Ziegler, 2020; Wright & Zimmermann, 2019; for exceptions see Tomko et al., 2014;

Zimmermann et al., 2019). Therefore, what is actually being measured in ambulatory assessment often remains hidden, despite a general increase of measurement precision in ambulatory assessment (Wright & Zimmermann, 2019).

Ambulatory assessment emerged as a method to specifically examine theoretical assumptions about person-situation processes. In contrast, other assessment methods either gradually incorporated situational components over time or person-situation theories were integrated or elaborated in hindsight to explain underlying psychological processes. The frame-of-reference in personality trait inventories is one example (Lievens, De Corte, et al., 2008; Shaffer & Postlethwaite, 2012). In this line of research, contextualized tags (e.g. at work) were added to items in personality questionnaires in order to increase the predictive validity. The frame-of-reference approach rests on the notion that individuals' tendencies for trait-relevant behavior may vary from context to context. Accordingly, providing a fixed contextual frame of reference that is matched to the context of the criterion leads to higher predictive validity of personality measures (e.g., job performance; Shaffer & Postlethwaite, 2012). A further example is research on assessment center exercises, which incorporated person-situation theories more consciously in recent years (Jansen et al., 2013; Lievens et al., 2006, 2009; Oliver et al., 2016). As assessment centers often lacked construct-related validity (e.g., Woehr & Arthur Jr, 2003) several studies built on Trait Activation Theory to develop approaches that help to increase the construct-relatedness of observations within and across exercises (e.g., Lievens et al., 2015; Oliver et al., 2016; Schollaert & Lievens, 2012). Furthermore, Jansen and colleagues (2013) demonstrated that the individual perception of situations influenced the participants' behaviors within assessment center exercises.

Situational Judgment Tests

SJTs share similar challenges to the above-mentioned assessment method with regard to the underlying processes. They were originally designed as simulations of relevant real-life situations (Motowidlo et al., 1990; Weekley et al., 2015). Although the history of these tests traces back much longer, Motowidlo and colleagues (1990) reintroduced SJTs to science and practice as useful tools for personnel selection. The development of SJT items typically relies on critical incidents that demonstrate effective behavior in terms of job performance (Corstjens et al., 2017). Therefore, the conceptual backbone of low-fidelity simulations is the assumption of behavioral consistency (Wernimont & Campbell, 1968). That is, behavior in simulated situations should predict behavior in similar real-life situations (Bruk-Lee et al., 2013; Lievens & De Soete, 2012). In line with this assumption, several meta-analyses confirmed a link between SJT responses and job performance (Christian et al., 2010; McDaniel et al., 2001, 2007). Although all SJTs focus on situation descriptions as common core, these tests reflect a methodological approach rather than a test of specific psychological processes (Lievens,

Peeters, et al., 2008). Hence, SJTs vary in form and design such as the response instruction and response format (Ployhart & Ehrhart, 2003). Mostly, participants are asked what they should or would do in a given situation (McDaniel & Nguyen, 2001). However, some SJTs also use open response formats (e.g., Rockstuhl et al., 2015) or participants are instructed to rate the effectiveness of response options rather than to choose a single response option (e.g., Motowidlo et al., 2006b). Furthermore, video-based situation descriptions are also common in addition to written situation descriptions (e.g., Lievens & Sackett, 2006).

Underlying Processes of Situational Judgment Tests. As SJTs are predominantly used for personnel selection, test developers were mostly concerned with the criterion-related validity of SJTs (Corstjens et al., 2017; Schmitt & Chan, 2006). However, a closer look at how SJTs function as psychological assessment tools reveals several caveats. Similar to assessment centers, SJTs often lack construct-related validity (Guenole et al., 2017; McDaniel et al., 2016). Most SJTs aggregate responses to several situation descriptions following the idea that these aggregated test-scores assess broad dimensions such as job skills or knowledge (see Bergman et al., 2006; Weekley et al., 2015). However, the true dimensionality of SJT scores is very seldomly assessed (Guenole et al., 2017). In fact, researchers frequently base conclusions about the underlying dimensionality of SJT scores on measures of internal consistency (Schmitt & Chan, 2006). As internal consistency is typically low (Catano et al., 2012; Kasten & Freund, 2016), SJTs are often described as multidimensional measures (Lievens, 2017b) that correlate with various constructs such as general mental ability and broad personality traits (McDaniel et al., 2007). This lack of understanding about psychological processes underlying SJT responses limits the interpretability of SJT performance and potentially attenuates correlations between SJT scores and relevant criteria (Wittmann & Klumb, 2006; see also Schulze et al., 2020).

To address this problem, construct-driven SJTs have been proposed (Guenole et al., 2017; Lievens, 2017b). Construct-driven SJTs are designed to measure an unidimensional construct (e.g., conscientiousness). Specifically, construct-driven SJTs build on Trait Activation Theory to develop situation descriptions based on trait-activating cues (Guenole et al., 2017; Lievens, 2017b). Moreover, all response options in construct-driven SJTs reflect behavior that represents different levels of the same unidimensional construct. In fact, response behavior in construct-driven SJTs is more consistent compared to traditional SJTs and test scores highly correlate to self-reports of the respective construct (Mussel et al., 2018; Olaru et al., 2019; Oostrom et al., 2018). Construct-driven SJTs are in line with Whole Trait Theory, as these tests rely on the assumption that the aggregates of several trait-related measures across situations relate to general personality traits.

In contrast to Whole Trait Theory, situation-specific variance is typically not

considered in construct-driven SJTs. One exception is a series of studies, in which consistent variance across situations as well as situation-specific variance in SJT responses was examined (Lievens et al., 2018). Notably, these authors found that aggregated SJT scores correlated with corresponding personality trait measures and that within-person variability in SJT responses correlated with variability in respective personality states. However, these correlations were only small to moderate. In order to examine situation-specific influences on SJT responses, other studies disentangled SJT responses into consistent variance across situations and situation-specific variance. However, the results were rather conflicting. That is, SJT responses were either overwhelmingly driven by individual SJT items (Westring et al., 2009) or almost no variance could be attributed to situation-specific processes (Jackson et al., 2016).

Krumm and colleagues (2015) took a more explicit approach to examine the relevance of situation descriptions for SJT responses. Across several studies, these authors applied SJT items with and without situation descriptions. Surprisingly, for the majority of items, item difficulty did not change when situation descriptions were omitted. Similar results were found for SJTs, in which situation descriptions consisted of short video sequences (Schäpers et al., 2020). Importantly, construct-related validity and the prediction of criteria differed only marginally between groups that responded to SJT items with and without situation descriptions (Schäpers et al., 2019). The authors concluded that SJTs may assess context-independent constructs rather than person-situation processes. Similarly, Rockstuhl et al. (2015) argued that the perception and judgment of situations in SJTs are not reflected in the responses to SJT items. When participants were asked separately what they would do in a given situation and how they perceive the situation, both responses correlated only moderately with each other and both responses predicted relevant criteria.

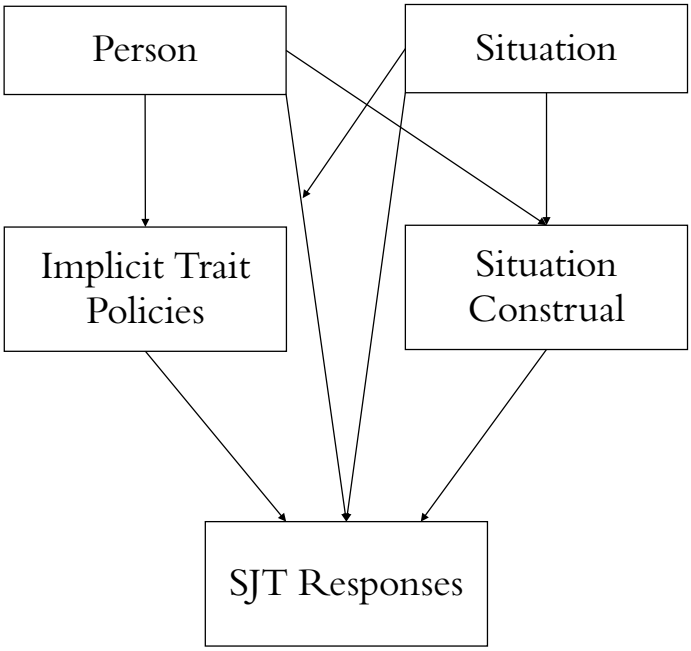
Based on these results, Lievens and Motowidlo (2016) argued that SJT responses rather reflect a context-independent construct, namely Implicit Trait Policies (ITPs). ITPs are defined as an individual's implicit belief about the effectiveness of trait-related behaviors (Lievens, 2017a; Motowidlo et al., 2006a, 2006b; Motowidlo & Beier, 2010). For example, some people may hold the belief that agreeable behavior is generally more effective than disagreeable behavior, regardless of the specific situation and context. Such implicit beliefs about the effectiveness of trait-related behaviors are thought to exist for any particular trait. Originally, ITPs were introduced to explain why SJTs correlate with personality traits, even when these traits were not intended to be assessed by the SJT (Motowidlo et al., 2006a, 2006b). Motowidlo and colleagues (2006a, 2006b) argued that individuals are more likely to believe in the effectiveness of behaviors, if individuals possess the personality trait the behavior reflects. In fact, several studies confirmed the link between personality traits and ITPs (Martin-Raugh et al., 2016; Motowidlo et al., 2006b; Oostrom et al., 2012). Empirical evidence for the notion that

ITPs are the underlying factors of SJT performance derives from a study that demonstrated a large overlap between SJT scoring keys developed by subject matter experts and novices (Motowidlo & Beier, 2010). The authors argued that novices do not possess situation-specific knowledge and experiences that enable them to select the most effective behavior in specific situations. Hence, novices must rely on general policies about the effectiveness of behaviors for their judgement (Motowidlo & Beier, 2010).

A Working Model of Situational Judgment Test Responses

Figure 2

Working Model of SJT Responses



Notes. The working model integrates the situation construal model (Funder, 2016; see also Schäpers et al., 2019 for a first adaption in the context of SJTs), Trait Activation Theory (Tett & Burnett, 2003; Tett & Guterman, 2000), and Implicit Trait Policies (Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010).

Lievens and Motowidlo (2016) reconceptualized SJTs as context-independent measures with an emphasis on ITPs. Although this perspective received support (e.g., Crook, 2016; Harvey, 2016; Krumm et al., 2015), several scholars argued that situations in SJT items may still be relevant to SJT performance (Brown et al., 2016; Fan et al., 2016; Harris et al., 2016; Melchers & Kleinmann, 2016). Since the wording of most response options contains situational cues, information presented in the response options may suffice to deduce the particular situation even if situation descriptions were

omitted in the SJT item (Fan et al., 2016; Harris et al., 2016; Melchers & Kleinmann, 2016). The debate about the underlying psychological processes revealed that SJT research typically fails to directly incorporate theories about person-situation processes (Brown et al., 2016; Harris et al., 2016; cf. Lievens et al., 2018; Schäpers et al., 2019). Especially Brown and colleagues (2016) pointed out that situational information in SJT items may be distinguished into situation cues, characteristics, and classes – similar to real-life situations. Likewise, Lievens (2017a) argued that individual responses to SJT items reflect real-life behavior in specific situations. Research on situation contingencies in behavior may specifically benefit from the use of SJTs (Lievens, 2017a). Since situations are identical across individuals, SJTs allow for direct comparisons of variance components in test-taker's responses that are contingent on the situation or consistent within individuals. Harris and colleagues (2016) posited that a specific case of situation contingencies may be the underlying factor of SJT performance. Following Trait Activation Theory, these authors argued that the strength of each SJT situation may influence to what degree a particular state corresponds to the individual's general tendency in behavior (i.e., trait). More precisely, the stronger the situation the less should individual responses be driven by personality traits. Figure 2 outlines a working model of underlying psychological processes of SJT responses that incorporates all previously proposed mechanisms. First, the model acknowledges direct effects of person characteristics such as general mental ability or personality traits on SJT responses. These effects are meta-analytically well established (McDaniel et al., 2007). Second, following the reconceptualization of SJTs as context-independent measures, the model specifically considers ITPs as relevant person characteristics for SJT responses (Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010). Third, the working model overcomes the vague definition of situational processes in SJTs as low-fidelity simulations by incorporating theoretical assumptions of person-situation processes. The working model is built particularly on the situation construal model (Funder, 2016; see also Schäpers et al., 2019). Thus, test-takers may construe psychologically relevant situations of SJT items based on their perception of situation cues in situation descriptions and response options of SJT items (see Brown et al., 2016). According to this reasoning, SJT responses reflect a psychological state for the given situation. Finally, the working model also includes core assumptions of Trait Activation Theory (Tett & Guterman, 2000): Trait activating cues in SJT items as well as situational strength of situations in SJT items may increase the correlation between personality traits and SJT responses (Guenole et al., 2017; Harris et al., 2016; Lievens, 2017b). This is depicted as a moderating effect of the situation on the relation of person characteristics and SJT responses in the working model.

In addition to the relation of person characteristics and SJT responses, theoretical arguments about possible processes underlying SJT responses are summarized in the working model. For the most part, thorough empirical studies examining these

processes are missing (c.f., Lievens et al., 2018; Motowidlo et al., 2018; Schäpers et al., 2019). This dissertation consists of seven studies that help uncovering the underlying psychological processes of SJTs (see Table 1 for an overview). In the following, I will briefly outline the scope of the subsequent chapters.

Situational Judgment Tests and Implicit Trait Policies

ITPs were originally conceptualized to explain why personality traits correlate with SJT performance even when these tests were applied with a knowledge instruction (i.e., asking test-takers what they should do; Motowidlo et al., 2006b). Accordingly, several studies exist that related ITPs to personality traits (Martin-Raugh et al., 2016; Motowidlo et al., 2006b, 2016, 2018; Oostrom et al., 2012). To assess ITPs, test-takers' effectiveness ratings of SJT response options should be correlated with the trait expression of these response options (see Lievens, 2017a; Lievens & Motowidlo, 2016). For instance, some response options of SJTs may reflect a high level of agreeableness whilst others reflect a low level of agreeableness. The correlation between the effectiveness rating and the trait expression of response options thus reflects how strong an individuals' rating is bound to the trait-level of response options. However, this operationalization aims to assess ITPs with the method they were designed to explain in the first place. Thus, the construct-related validity of ITP measures is of particular interest in order to gain meaningful insights about the relevance of ITPs for SJT responses. Chapter 2 describes two studies that examine the construct-related validity of ITPs.

Situational Judgment Tests and Person-Situation Processes

To assess the influence of situation descriptions for SJT responses, previous research relied on the manipulation of situation cues in SJT items (Krumm et al., 2015; Schäpers et al., 2019, 2020; c.f., Lievens et al., 2018; Rockstuhl et al., 2015). Although these results yielded valuable insights about underlying processes of SJT responses, those studies did not utilize insights on person-situation processes to their fullest potential. Building on research about real-life person-situation processes, the working model of SJT responses (Figure 2) incorporates the situation construal as an underlying factor of SJT responses. Chapter 3 examines whether an individual's situation construal of SJT items affects responses to the same item. Specifically, the three presented studies incorporate a recent taxonomy of situation characteristics (Rauthmann et al., 2014) in order to explicitly assess the situation construal of SJT items. Three questions are of particular interest: (a) does the situation construal of SJT items predict SJT responses, (b) which test elements (i.e., situation descriptions or response options) of SJT items evoke a relevant situation construal, and (c) does the situation construal of SJT items predict relevant criteria?

Beyond situation construal, Trait Activation Theory makes specific assumptions

how situation cues influence the relation of personality traits and behavior (Tett & Guterman, 2000). Building on Trait Activation Theory, arguments were brought forward that trait-activating cues in situation descriptions of construct-driven SJTs increase the relevance of personality traits for response behaviors (Guenole et al., 2017; Lievens, 2017b) and that the strength of situation cues in SJT items further moderates the relation of personality and SJT responses (Harris et al., 2016). However, neither claims have been tested. A study presented in Appendix C tests whether omitting trait-activating cues from SJT items reduce the construct-related validity of an SJT assessing narrow personality facets. Further, Chapter 4 examines the influence of SJT items´ situational strength on the relation of personality and SJT responses. Overall, these studies represent a theory-driven investigation of person-situation processes that may underly SJT responses.

Standardized State Assessment

So far, studies included in this dissertation were concerned with the underlying processes of SJTs, specifically, whether person-situation processes are relevant for response behavior in SJT items. However, previous studies revealed that SJTs have major limitations as a methodological approach. Noteworthy are the lack of construct-related validity as well as the lack of psychological relevance of essential test elements (i.e., situation descriptions) for response behavior (see Krumm et al., 2015; McDaniel et al., 2016). Thus, Chapter 5 takes a step back to deduce how person-situation processes may be assessed with situation descriptions. Whole Trait Theory serves as theoretical base to propose a methodological framework for the assessment of psychological states in hypothetical situations – Standardized State Assessment. To do so, I take a closer look at how real-life states are assessed and how these principles may be applied to Standardized State Assessment. Moreover, I outline how situation descriptions may be developed to increase the psychological similarity to real-life situations. Building on the research presented in this dissertation, Chapter 5 further contrasts SJTs and Standardized State Assessment and concludes with methodological guidelines for researchers interested in assessing person-situation processes.

Summary

Naturally, theoretical advances on person-situation processes sparked an increase in psychological assessments that considered these processes. SJTs are no exceptions to this development. However, research did not find a consensus about what psychological processes underly SJT performance. Arguments have been brought forward that support person-situation processes but also completely new and context-independent constructs (i.e., ITPs). However, the empirical support of either processes is rather limited.

So far, I introduced a working model of SJT responses to summarize all existing propositions about underlying processes of SJT performance. The following chapters add to the empirical knowledge about these processes and thus contribute to resolving the debate whether SJTs reflect measures of person-situation processes or context-independent measures. Finally, this dissertation will conclude by proposing a methodological framework that adds to the core of contemporary personality research: the assessment of person-situation processes. Using situation descriptions to assess these processes would undoubtedly be beneficial for various research questions on person-situation interactions. This method especially has the potential for much more economic assessments when compared to ambulatory assessments or assessment center exercises and even allows researchers to sample uncommon situations. In sum, this framework suggests how this goal may be achieved by building on theory and best practices of SJT development

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*(2), 150–166. https://doi.org/10.1177/1088868306294907

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, J. J. A., Fleeson, W., Grafton, B., Jayawickreme, E., Kurzius, E., MacLeod, C., Miller, L. C., Read, S. J., Roberts, B., Robinson, M. D., Wood, D., Wrzus, C., & Mõttus, R. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality*, *31*(5), 503–528. https://doi.org/10.1002/per.2115

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*(3), 223–235. https://doi.org/10.1111/j.1468-2389.2006.00345.x

Bledow, R., & Frese, M. (2009). Situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, *62*(2), 229–258. https://doi.org/10.1111/j.1744-6570.2009.01137.x

Bleidorn, W. (2009). Linking personality states, current social roles and major life goals. *European Journal of Personality*, *23*(6), 509–530. https://doi.org/10.1002/per.731

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*(1), 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). Springer.

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*(3), 333–346. https://doi.org/10.1111/j.1468-2389.2012.00604.x

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*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgment tests for selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention* (pp. 228–248). John Wiley & Sons Ltd.

Crook, A. E. (2016). Unintended consequences: Narrowing SJT usage and losing credibility with

applicants. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *9*(1), 59–63. https://doi.org/10.1017/iop.2015.118

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Dweck, C. S. (2017). From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, *124*(6), 689–719. https://doi.org/10.1037/rev0000082

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*(7), 1097–1126. https://doi.org/10.1037/0022-3514.37.7.1097

Epstein, S., & O'Brien, E. J. (1985). The person–situation debate in historical and current perspective. *Psychological Bulletin*, *98*(3), 513–537. https://doi.org/10.1037/0033-2909.98.3.513

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*(1), 43–47. https://doi.org/10.1017/iop.2015.114

Fisher, A. J., & Boswell, J. F. (2016). Enhancing the personalization of psychotherapy with dynamic assessment and modeling. *Assessment*, *23*(4), 496–506. https://doi.org/10.1177/1073191116638735

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior*, *33*(7), 865–877. https://doi.org/10.1002/job.1803

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009

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*(4), 1667–1684. https://doi.org/10.1111/j.1751-9004.2008.00122.x

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, *40*(1), 21–34. https://doi.org/10.1016/j.jrp.2005.08.003

Funder, D. C. (2016). Taking situations seriously: The situation construal model and the Riverside Situational Q-Sort. *Current Directions in Psychological Science*, *25*(3), 203–208. https://doi.org/10.1177/0963721416635552

Geiser, C., Hintz, F., Burns, G. L., & Servera, M. (2017). Latent variable modeling of person-situation data. In J. F. Rauthmann, R. A. Sherman, & D. C. Funder (Eds.), *The Oxford handbook of psychological situations*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190263348.013.15

Guenole, N., Chernyshenko, O. S., & Weekly, J. A. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, *17*(3), 234–252. https://doi.org/10.1080/15305058.2017.1297817

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology*, *55*(1), 137–163. https://doi.org/10.1111/j.1744-6570.2002.tb00106.x

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

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*(1), 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*(1), 63–71. https://doi.org/10.1017/iop.2015.119

Hofmans, J., De Clercq, B., Kuppens, P., Verbeke, L., & Widiger, T. A. (2019). Testing the structure and process of personality using ambulatory assessment data: An overview of within-person and person-specific techniques. *Psychological Assessment*, *31*(4), 432–443. https://doi.org/10.1037/pas0000562

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). SAGE.

Horstmann, K. T., & Ziegler, M. (2020). Assessing personality states: What to consider when constructing personality state measures. *European Journal of Personality*. https://doi.org/10.1002/per.2266

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2016). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology*, *90*(1), 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*(2), 326–341. https://doi.org/10.1037/a0031257

Jayawickreme, E., Zachry, C. E., & Fleeson, W. (2019). Whole trait theory: An integrative approach to examining personality structure and process. *Personality and Individual Differences*, *136*, 2–11. https://doi.org/10.1016/j.paid.2018.06.045

Jones, A. B., Brown, N. A., Serfass, D. G., & Sherman, R. A. (2017). Personality and density distributions of behavior, emotions, and situations. *Journal of Research in Personality*, *69*, 225–236. https://doi.org/10.1016/j.jrp.2016.10.006

Kantor, J. R. (1924). *Principles of psychology*. Knopf.

Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). *European Journal of Psychological Assessment*, *32*(3), 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*(2), 399–417. https://doi.org/10.1037/a0037674

Lewin, K. (1936). *Principles of topological psychology*. McGraw-Hill.

Lievens, F. (2017a). Assessing personality-situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*(5), 424–440. https://doi.org/10.1002/per.2111

Lievens, F. (2017b). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing*, *17*(3), 269–276. https://doi.org/10.1080/15305058.2017.1309857

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology*, *91*(2), 247–258. https://doi.org/10.1037/0021-9010.91.2.247

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*(8), 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 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*(7), 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*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*(5), 1181–1188. https://doi.org/10.1037/0021-9010.91.5.1181

Lievens, F., Schäpers, P., & Herde, C. N. (2020). Situational judgment tests: From low-fidelity simulations to alternative measures of personality and the person-situation interplay. In D. Wood, P. Harms, S. Read, & A. Slaughter (Eds.), *Emerging approaches to measuring and modeling the person and*

situation. Elsevier.
Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology*, *100*(4), 1169–1188. https://doi.org/10.1037/apl0000004
Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao (Eds.), *Research in personnel and human resources management* (pp. 99–152). Emerald Group Publishing Limited. https://doi.org/10.1108/S0742-7301(2009)0000028006
Martin-Raugh, M. P., Kell, H. J., & Motowidlo, S. J. (2016). Prosocial knowledge mediates effects of agreeableness and emotional intelligence on prosocial behavior. *Personality and Individual Differences*, *90*, 41–49. https://doi.org/10.1016/j.paid.2015.10.024
McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81
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*(1), 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: Perspectives on Science and Practice*, *9*(1), 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*(4), 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*(1-2), 103–113. https://doi.org/10.1111/1468-2389.00167
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*(1), 29–34. https://doi.org/10.1017/iop.2015.111
Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*(1), 121–140. https://doi.org/10.1177/0149206309349309
Meyer, R. D., Dalal, R. S., José, I. J., Hermida, R., Chen, T. R., Vega, R. P., Brooks, C. K., & Khare, V. P. (2014). Measuring job-related situational strength and assessing its interactive effects with personality on voluntary work behavior. *Journal of Management*, *40*(4), 1010–1041. https://doi.org/10.1177/0149206311425613
Mischel, W. (1968). *Personality and assessment*. Psychology Press.
Mischel, W. (1977). The interaction of person and situation. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333–352). Lawrence Erlbaum.
Mischel, W. (1979). On the interface of cognition and personality: Beyond the person-situation debate. *American Psychologist*, *34*(9), 740–754. https://doi.org/10.1037/0003-066X.34.9.740
Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. 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*(2), 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*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640
Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance*, *29*(4), 331–346. https://doi.org/10.1080/08959285.2016.1165227
Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). 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). Lawrence Erlbaum Associates.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*(4), 749–761. https://doi.org/10.1037/0021-9010.91.4.749

Motowidlo, S. J., Lievens, F., & Ghosh, K. (2018). Prosocial implicit trait policies underlie performance on different situational judgment tests with interpersonal content. *Human Performance*, *31*(4), 238–254. https://doi.org/10.1080/08959285.2018.1523909

Mussel, P., Gatzka, T., & Hewig, J. (2018). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment*, *34*(5), 328–335. https://doi.org/10.1027/1015-5759/a000346

Nestler, S., Humberg, S., & Schönbrodt, F. D. (2019). Response surface analysis with multilevel data: Illustration for the case of congruence hypotheses. *Psychological Methods*, *24*(3), 291–308. https://doi.org/10.1037/met0000199

Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research. *Journal of Personnel Psychology*, *9*(2), 79–93. https://doi.org/10.1027/1866-5888/a000009

Olaru, G., Burrus, J., Maccann, C., Zaromb, M. F., Wilhelm, O., & Roberts, D. R. (2019). Situational judgment tests as a method for measuring personality: Development and validity evidence for a test of dependability. *PLoS One*, *14*(2), e0211884. https://doi.org/10.1371/journal.pone.0211884

Oliver, T., Hausdorf, P., Lievens, F., & Conlon, P. (2016). Interpersonal dynamics in assessment center exercises: Effects of role player portrayed disposition. *Journal of Management*, *42*(7), 1992–2017. https://doi.org/10.1177/0149206314525207

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*(4), 335–353. https://doi.org/10.1080/08959285.2012.703732

Oostrom, J. K., de Vries, R. E., & de Wit, M. (2018). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*(1), 1–29. https://doi.org/10.1080/08959285.2018.1539856

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2016). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology*, *112*(4), 642–681. https://doi.org/10.1037/pspp0000111

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), 1–16. https://doi.org/10.1111/1468-2389.00222

Rauthmann, J. F. (2015). Structuring situational information. A road map of the multiple pathways to different situational taxonomies. *European Psychologist*, *20*(3), 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., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677–718. https://doi.org/10.1037/a0037250

Rauthmann, J. F., Jones, A. B., & Sherman, R. A. (2016). Directionality of person-situation transactions: Are there spillovers among and between situation experiences and personality states? *Personality and Social Psychology Bulletin*, *42*(7), 893–909. https://doi.org/10.1177/0146167216647360

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*(3), 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*(4), 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*(2), 464–480. https://doi.org/10.1037/a0038098

Sackett, P. R., & Lievens, F. (2008). Personnel Selection. *Annual Review of Psychology*, *59*(1), 419–450. https://doi.org/10.1146/annurev.psych.59.103006.093716

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*(3), 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. (2020). 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*, *93*(2), 472–494. https://doi.org/10.1111/joop.12297

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

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

Schollaert, E., & Lievens, F. (2012). Building situational stimuli in assessment center exercises: Do specific exercise instructions and role-player prompts increase the observability of behavior? *Human Performance*, *25*(3), 255–271. https://doi.org/10.1080/08959285.2012.683907

Schulze, J., West, S. G., Freudenstein, J.-P., Schäpers, P., Mussel, P., Eid, M., & Krumm, S. (2020). *Hidden framings and hidden asymmetries in the measurement of personality – A combined lens-model and frame-of-reference perspective* [Manuscript under review].

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*(3), 445–494. https://doi.org/10.1111/j.1744-6570.2012.01250.x

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*(5), 872–888. https://doi.org/10.1037/pspp0000036

Sonnentag, S., & Binnewies, C. (2013). Daily affect spillover from work to home: Detachment from work and sleep as moderators. *Journal of Vocational Behavior*, *83*(2), 198–208. https://doi.org/10.1016/j.jvb.2013.03.008

Steyer, R., Ferring, D., & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*(2), 79–98.

Steyer, R., & Schmitt, M. (1990). The effects of aggregation across and within occasions on consistency, specificity and reliability. *Methodika*, *4*, 58–94.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Tett, & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 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*(4), 397–423. https://doi.org/10.1006/jrpe.2000.2292

Thornton III, G. C., & Rupp, D. E. (2004). Simulations and assessment centers. In J. C. Thomas (Ed.), *Comprehensive handbook of psychological assessment, Vol. 4: Industrial and organizational assessment* (pp. 319–344). John Wiley & Sons Inc.

Tomko, R. L., Solhan, M. B., Carpenter, R. W., Brown, W. C., Jahng, S., Wood, P. K., & Trull, T. J. (2014). Measuring impulsivity in daily life: The momentary impulsivity scale. *Psychological Assessment*, *26*(2), 339–349. https://doi.org/10.1037/a0035083

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, *23*(6), 466–470. https://doi.org/10.1177/0963721414550706

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual*

Review of Organizational Psychology and Organizational Behavior, *2*(1), 295–322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

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). Lawrence Erlbaum Associates.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, *52*(5), 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*(1), 44–63. https://doi.org/10.1080/08959280802540999

Wilson, R. E., Thompson, R. J., & Vazire, S. (2017). Are fluctuations in personality states more than fluctuations in affect? *Journal of Research in Personality*, *69*, 110–123. https://doi.org/10.1016/j.jrp.2016.06.006

Wittmann, W. W., & Klumb, P. L. (2006). How to fool yourself with experiments in testing theories in psychological research. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 185–211). American Psychological Association.

Woehr, D. J., & Arthur Jr, W. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management*, *29*(2), 231–258.

Wright, A. G., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment*, *31*(12), 1467–1480. https://doi.org/10.1037/pas0000685

Zimmermann, J., Woods, W. C., Ritter, S., Happel, M., Masuhr, O., Jaeger, U., Spitzer, C., & Wright, A. G. C. (2019). Integrating structure and dynamics in personality assessment: First steps toward the development and validation of a personality dynamics diary. *Psychological Assessment*, *31*(4), 516–531. https://doi.org/10.1037/pas0000625

Chapter 2

On the Construct-Related Validity of Implicit Trait Policies

This article has been prepared for publication:

Freudenstein, J.-P., Mussel, P., & Krumm, S. (2020). *On the construct-related validity of implicit trait policies* [Manuscript prepared for publication].

Chapter 2

On the Construct-Related Validity of Implicit Trait Policies

Jan-Philipp Freudenstein, Patrick Mussel, & Stefan Krumm
Freie Universität Berlin

In response to recent calls to incorporate Implicit Trait Policies (ITPs) into personality research, the current research examined the construct-related validity of ITP measures in two studies. ITPs are defined as implicit beliefs about the effectiveness of behaviors that reflect a certain trait. They are assessed by utilizing the methodology of Situational Judgment Tests. In the first study, we employed Monte Carlo Simulation to highlight possible caveats when interpreting correlations between ITP scores and SJT scores that are derived from the same test. In the second study, we empirically examined (N = 339) several underlying key assumptions of ITP theory, including trait-specificity, the relation to personality traits, their context-independence, and the relation to general domain knowledge. Overall, our results showed little support for these assumptions. Although we found some confirmation for expected correlations between ITPs and personality traits, most of the observed variance in ITP measures was either method specific or due to measurement error. We conclude that ITP measures lack construct-related validity and discuss implications for SJT theory and beyond.

Keywords. Implicit Trait Policies, Situational Judgment Tests, Validity

Building on the notion that both person and situation drive individual behavior, a considerable amount of research focused on dynamic processes that explain these relations (e.g., Baumert et al., 2017; Fleeson, 2007; Funder, 2016; Mischel & Shoda, 1995; Rauthmann et al., 2014). Recently, Implicit Trait Policies (ITPs) have been proposed as a situation-independent construct that might mediate the link between traits and behavior, and thus "enhance contemporary personality research" (Lievens, 2017a, p. 431; see also Martin-Raugh et al., 2016). ITPs are defined as implicit beliefs about the effectiveness of behaviors that reflect a certain trait (Motowidlo et al., 2006b). For instance, the belief that agreeable behavior is generally more effective than disagreeable behavior constitutes an ITP for agreeableness; the belief that extraverted behavior is generally more effective than introverted behavior describes an ITP for extraversion (Lievens & Motowidlo,

Correspondence concerning this paper should be addressed to Jan-Philipp Freudenstein, Institute of Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany. E-Mail: jan-philipp.freudenstein@fu-berlin.de

We thank Stephan J. Motowidlo and Filip Lievens for helpful comments on an earlier version of this manuscript.

All data and R code are available on the Open Science Framework (osf.io/ pcu3j).

2016). That is, ITPs reveal an individual's effectiveness rating of specific behaviors, which can be attributed to the trait-level expression of the behavior (Lievens, 2017a). Although the concept of ITPs is currently closely related to the method and theory of Situational Judgment Tests (SJTs; Lievens & Motowidlo, 2016), several researchers agreed that ITPs may be fruitful for personality research in general (e.g., Judge et al., 2017; Lievens, 2017a; Motowidlo, 2017; Wright, 2017).

Despite ITPs' relevance for SJT research and beyond, our knowledge about ITP measurement and its validity is still sparse. Since ITPs are (so far) exclusively assessed with SJTs, their construct-related validity may be of particular concern. That is, the construct-related validity of SJTs was recently described as "hot mess" (McDaniel et al., 2016, p. 47). Being closely tied to SJTs, ITPs may share the same fate. The currently available evidence on the construct-related validity of ITPs is limited, as most recent studies did not simultaneously assess ITPs for different traits and across several SJTs, thus precluding insights on their convergent and discriminant construct-related validity (e.g., Motowidlo et al., 2006b, 2018).

Construct-related validity is a necessary condition to derive meaningful conclusions about the relevance of ITPs for SJT responses or individual behavior in general (see Cronbach & Meehl, 1955). In this study, we address the measurement quality and construct-related validity of ITPs to facilitate future research on this concept. First, we utilize Monte Carlo simulation to highlight possible caveats when interpreting correlations between ITP scores and SJT scores that are derived from the same test. Second, we scrutinize the construct-related validity of ITPs when measured with several SJTs. In doing so, we contribute to a deeper understanding of the measurement quality of ITPs. This is essential to ensure high quality research when adopting the concept of ITPs to explain SJT functioning and, more broadly, when using ITPs in personality research.

Implicit Trait Policies

The theoretical foundation of ITPs rests on the assumption of dispositional fit (Motowidlo et al., 2006a, 2006b). Accordingly, people develop implicit beliefs about the effectiveness of specific behaviors across the lifespan, which are in line with the personality traits individuals possess (e.g., agreeable people are more likely to believe that agreeable behavior is more effective; Motowidlo et al., 2006a, 2006b). Motowidlo (2003) used this argument to explain the relation of personality traits and job performance. For instance, "when a problematic work situation demands an expression of a particular trait for effective resolution, people who possess that trait are more likely to believe that behaviors expressing that trait will be effective in that situation" (Motowidlo et al., 2006b, p. 751). Thus, an ITP for a specific trait is conceptualized as a causal link between that personality trait and trait-related behavior. So, individuals high on agreeableness will act agreeable because they hold the belief, or the ITP, that agreeable behavior is effective. Although being influenced by individuals' personality dispositions, ITPs can develop through general life experiences (Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo & Peterson, 2008). For instance, Motowidlo and Peterson (2008) showed that prison inmates had stronger ITPs for agreeableness when compared to correctional officers. The authors

concluded that differences in organizational perspective may influence ITPs.

The relation between self-reports of a personality trait and the ITP for that trait has been examined across several trait domains (Martin-Raugh et al., 2016; Motowidlo et al., 2006b, 2016, 2018; Oostrom et al., 2012). As expected, correlations between self-reports of personality traits and corresponding ITPs were moderate (e.g., *r*s = .15 - .39; Motowidlo et al., 2006b). ITPs were also shown to predict behavior in role-play scenarios (Martin-Raugh et al., 2016; Motowidlo et al., 2006b)—although this was only true for agreeableness ITPs and not for extraversion ITPs. The authors attributed this mixed finding to shortcomings in the assessment of behavior (Motowidlo et al., 2006b). Motowidlo and Beier (2010) also revealed that both agreeableness ITPs and conscientiousness ITPs were moderately related to supervisor ratings of job performance.

Implicit Trait Policies in Situational Judgment Test Theory

ITPs have been introduced as an integral part of SJT theory (Lievens & Motowidlo, 2016; Motowidlo et al., 2006a, 2006b). SJTs are standardized tests that consist of short situation descriptions with several behavioral response options (McDaniel & Nguyen, 2001). They are popular tools in personnel selection mostly due to their predictive validity of job performance (e.g., Christian et al., 2010; McDaniel et al., 2001). Importantly, the link between SJT performance and job performance rests on the assumption that the processes of responding to SJTs are similar to real life behavior (Lievens & De Soete, 2012; Motowidlo et al., 1990; Weekley et al., 2015). Accordingly, SJTs are also often described as low-fidelity simulations (Motowidlo et al., 1990; Weekley et al., 2015).

Despite this popular view on SJTs, a growing body of research exists that challenges the notion of SJTs as simulations (Jackson et al., 2016; Krumm et al., 2015; Schäpers et al., 2019; Schäpers, Lievens, et al., 2020). For instance, Krumm et al. demonstrated that omitting situation descriptions from SJT items did not change item difficulty for a majority of all tested items. This finding has been further extended to video-based SJTs (Schäpers, Lievens, et al., 2020). Even for SJT items with highly specific video sequences as situation descriptions, the decrease in item difficulty when leaving out situation descriptions was the same as in text-based SJTs. Moreover, situation descriptions of SJT items had only negligible effects on SJT's construct and criterion-related validity (Schäpers et al., 2019; Schäpers, Freudenstein, et al., 2020). In sum, a considerable amount of evidence supports the view that situation descriptions—the part of an SJT item that is thought to "simulate" reality—are less relevant for SJT response behavior.

Lievens and Motowidlo (2016) suggested that the concept of ITPs may explain why some SJTs "worked" even without situation descriptions (see also Krumm et al., 2015; Motowidlo et al., 2006a, 2006b). They argued that test-takers may not only rely on their situation-specific experiences and knowledge to come up with a response to a fictional situation (as described in an SJT item). Test-takers may also rely on their ITPs, i.e., their general beliefs about the effectiveness of trait-related behavior described in response options. Since general beliefs about the effectiveness of a trait are not tied to a specific context, test-takers may rely on ITPs regardless of the specific

situation presented in an SJT item. Evidence in favor of ITPs as part of SJT responding is provided by Motowidlo and Beier (2010). These authors showed that SJT scoring keys from novices and subject-matter expert ratings (SMEs) largely converged. The authors concluded that novice raters had no relevant job knowledge and thus had to utilize general beliefs about effective behavior (i.e., ITPs) to construe a scoring key similar to the one created by experts.

Assessment of Implicit Trait Policies

Currently, ITPs are only *indirectly* assessed through individuals' responses to SJT items (Motowidlo et al., 2006b). In SJTs, individuals typically rate the effectiveness of behavioral responses to fictional situation descriptions (e.g., McDaniel & Nguyen, 2001; Weekley et al., 2015). SJT scores reflect the extent to which an individual's effectiveness rating corresponds with the actual effectiveness of a behavior (as, for example, determined by experts). ITP scores, on the contrary, specify the relation of an individual's effectiveness rating with the trait relatedness of a specific behavioral response (e.g., Lievens & Motowidlo, 2016). So, if an individual shows a tendency to rate responses reflecting extraverted behavior as effective and introverted behavior as ineffective, this individual would show a strong ITP for extraversion (see Table 1 for an example; Lievens, 2017a; Lievens & Motowidlo, 2016; Motowidlo et al., 2006b). Thus, the correlation between test-takers' effectiveness ratings of SJT responses and trait relatedness of each response (as determined by experts) is used as the ITP score.

As mentioned above, SJT research has been concerned with the question of how relevant ITPs are for SJT responding. Note that both, SJT scores and ITP scores are derived by comparing the same test-takers response with different scoring keys (see Table 1 for examples of such keys). In the case of SJT scores, this key is typically derived from SMEs' ratings of the effectiveness of the response options. In the case of ITP scores, the scoring key is based on SME ratings of the trait relatedness of response options. Typically, the extent to which these two keys are similar (i.e., correlated) is viewed as an SJT's saturation with ITPs (Motowidlo et al., 2018; see Table 1). In other words, the ITP saturation of an SJT defines the extent to which responding solely on the basis of response options' trait relatedness also results in effective responding.

It is not uncommon for SJTs that the same responses are scored in two or more ways (i.e., with two or more keys; e.g., Bergman et al., 2006; Ployhart & Ehrhart, 2003). However, it is problematic to interpret the correlation between two scores that constitute a computational variation of the same individual response. This is the case when empirical SJT scores are correlated with empirical ITP scores (e.g., Motowidlo et al., 2018; Motowidlo & Beier, 2010; Oostrom et al., 2012). The problem is similar to Pearson's notion of spurious correlations between ratio variables with the same denominator (Pearson, 1897). Spurious correlations are correlations that occur due to a shared denominator of two variables (e.g., population size) even though the variables are otherwise uncorrelated (Kronmal, 1993; Pearson, 1897). Similarly, when correlating SJT and ITP scores for the same set of responses, both scores share the individuals' identical responses. Thus, the unique variance in SJT and ITP scores is determined by the SJT and ITP scoring keys.

Table 1

Example of Scoring ITPs and Effectiveness of SJT responses

		Scoring Keys			Responses	
		Effectiveness Level of Response Options	**Trait Level of Response Options**		**Effectiveness Rating Person A**	**Effectiveness Rating Person B**
SJT Item 1	Response option 1	1	1		2	2
	Response option 2	5	5		5	5
	Response option 3	2	1		1	3
	Response option 4	3	1		3	3
SJT Item 2	Response option 5	2	5		4	2
	Response option 6	5	3		4	3
	Response option 7	3	5		5	3
	Response option 8	2	3		3	1
		Score Saturation: *r* = .42		**Effectiveness Score:**	*r* = .65	*r* = .86
				ITP Score:	*r* = .88	*r* = .25

Notes. This table contains two exemplary SJT items with four response options each and responses by two individuals. The effectiveness level (i.e., how effective is the behavior in the given situation) and trait level (e.g., how representative is the behavior for agreeableness) of all response options have been rated by subject matter experts. These ratings reflect the two scoring keys. The correlation of the two scoring keys represents the score saturation. To compute effectiveness and ITP scores, responses of each individual are correlated with the respective scoring key.

To illustrate this, we conducted a simulation study (Study 1). Specifically, we aimed to identify the extent to which the correlation of SJT and ITP scores for the same test responses are confounded as a function of the SJT's ITP saturation.

RQ1: To what extent are correlations of SJT and ITP scores of the same test responses confounded as a function of the SJT's ITP saturation?

Beyond the use of existing SJTs for the assessment of ITPs, Motowidlo et al. (2006b) also suggested to develop SJTs specifically for the assessment of ITPs. In such SJTs, all response options are created to reflect high or low levels of a specific trait. Additionally, response options that reflect high trait levels are always effective in the given situation and response options that reflect low trait levels are always ineffective. That is, the SJT score is completely saturated with a specific trait.

As mentioned before, evidence scrutinizing the validity of these methods is scarce (cf. Motowidlo et al., 2006b, 2018). However, the theory behind ITPs makes several assumptions which can be used to delineate the nomological net of ITPs. First, ITPs are defined as being trait-specific (Motowidlo et al., 2006b, 2018). That is, the implicit belief about the effectiveness of behaviors is bound to the trait these behaviors reflect. In fact, Motowidlo et al (2006b) showed that scores for ITPs of different traits (i.e., extraversion, agreeableness, conscientiousness) showed no substantial correlations among each other.

Second, ITPs are related to but distinct from personality traits. Building on research on dispositional fit, personality traits should help to develop specific ITPs (i.e., agreeable people are much more likely to belief that agreeable behavior is effective; Motowidlo et al., 2006b). However, ITPs are conceptualized as different constructs than personality traits (Lievens, 2017a; Motowidlo et al., 2006b; Motowidlo & Beier, 2010). Indeed, several studies revealed only small to medium correlations between (self-reported) personality traits and corresponding ITPs (Martin-Raugh et al., 2016; Motowidlo et al., 2006b, 2016, 2018).

Third, ITPs are defined as general, context independent constructs (Lievens & Motowidlo, 2016). Thus, measures of ITPs should not depend on the situational context presented in a specific SJT, but rather generalize across several SJTs. To examine this, Motowidlo, Lievens, and Gosh (2018) assessed prosociality ITPs with four different SJTs. Contrary to the notion of ITPs' context-independency, correlations among ITP scores ranged from $r = .22$ to $r = .46$.

Fourth, previous research has put forward the notion that ITPs reflect general domain knowledge (i.e., knowledge acquired through general experience; Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010). In their focal article, Lievens and Motowidlo (2016) explained this for agreeableness: "If it is true that behavior that expresses some personality trait such as agreeableness […] contributes to effective job performance, people who believe this have more general domain knowledge" (p. 9). In other words, this view posits that when ITPs "are accurate, they represent general domain knowledge because accurate ITPs can be learned before people enter any particular job situation and are not dependent on specific job experience" (Motowidlo & Beier, 2010, p. 323). In those instances, ITPs reflect justified true beliefs as an individual's ITP turns out to be true and

was justified by previous life experiences (e.g., a person experiences that agreeable behavior is generally effective and develops and ITP for agreeableness). However, justified true beliefs may not always reflect knowledge as the principles justifying the belief may not be accurate (see Gettier, 1963); that is, trait-related behaviors vary in their true effectiveness among situations. For instance, agreeable behavior will be effective to solve certain team conflicts but will be rather ineffective to achieve the best result in certain negotiations. Thus, ITPs may not be seen as knowledge as the belief itself does not express knowledge about when and why trait-related behaviors are effective. Hence, the question arises whether ITPs reflect a general belief or general domain knowledge. SJTs that were specifically designed to assess ITPs do not allow to examine this question. These tests are characterized by a perfect saturation of the response options' true effectiveness and the trait expressiveness. Thus, the knowledge about the effectiveness of behavior in a specific-situation is used as indicator for the relevant ITP. It follows that measures of ITPs can only differentiate between general domain knowledge and a general belief if the score saturation of these measures is imperfect (i.e., the true effective response options vary in their trait-relatedness). These measures would assess general domain knowledge if only response options for which the true effectiveness and the trait expression align load on a common factor. If all response options load on a common factor, these measures would reflect a general belief about the effectiveness of behaviors that reflect a certain trait.

To our knowledge no previous study exists that examined all of the above-mentioned core assumptions of ITP theory simultaneously. Thus, at this point no definite conclusions about the intended interpretability of these scores (i.e., as ITPs for agreeableness) can be drawn. Importantly, such inferences are essential to establish construct-related validity of specific measurements (Cronbach & Meehl, 1955). Especially, since psychological constructs are not directly observable (i.e., latent), a measure's construct-related validity should be examined in the context of relations to other methods and constructs (Campbell & Fiske, 1959). Following the outline of ITP theory, we expect that ITPs for the same trait are highly correlated even when measured with different SJTs. We expect ITPs for the same trait to correlate higher with each other than with their corresponding personality trait. We also expect ITPs for different traits to show only small correlations.

H1a: Different SJTs that measure ITPs for the same personality trait are significantly correlated (convergent validity).

H1b: Convergent correlations of ITPs are higher than the correlations of personality traits and ITPs of the same trait.

H2: Correlations of SJTs that measure ITPs for different personality traits are lower than their convergent correlations (discriminant validity).

Regarding the definition of ITPs as general domain knowledge, we phrased an open research question.

RQ2: Do measures of ITPs reflect general domain knowledge?

Study Overview

In two studies we aim to shed light on the construct-related validity of ITP measures. First, we use Monte Carlo simulation to investigate possible pitfalls

when interpreting correlations of ITP and SJT scores that were derived from the same set of responses (RQ1). Second, we gathered empirical data to assess convergent and discriminant relations of ITPs that were assessed with several methods (H1a – H2). This data will also help understanding the link between ITPs and general domain knowledge (RQ2). We preregistered Study 2, including all related hypotheses and research questions, on the Open Science Framework (osf.io/ m5ce8). Data and code are available for both studies on the Open Science Framework (osf.io/ pcu3j).

Study 1

Procedure

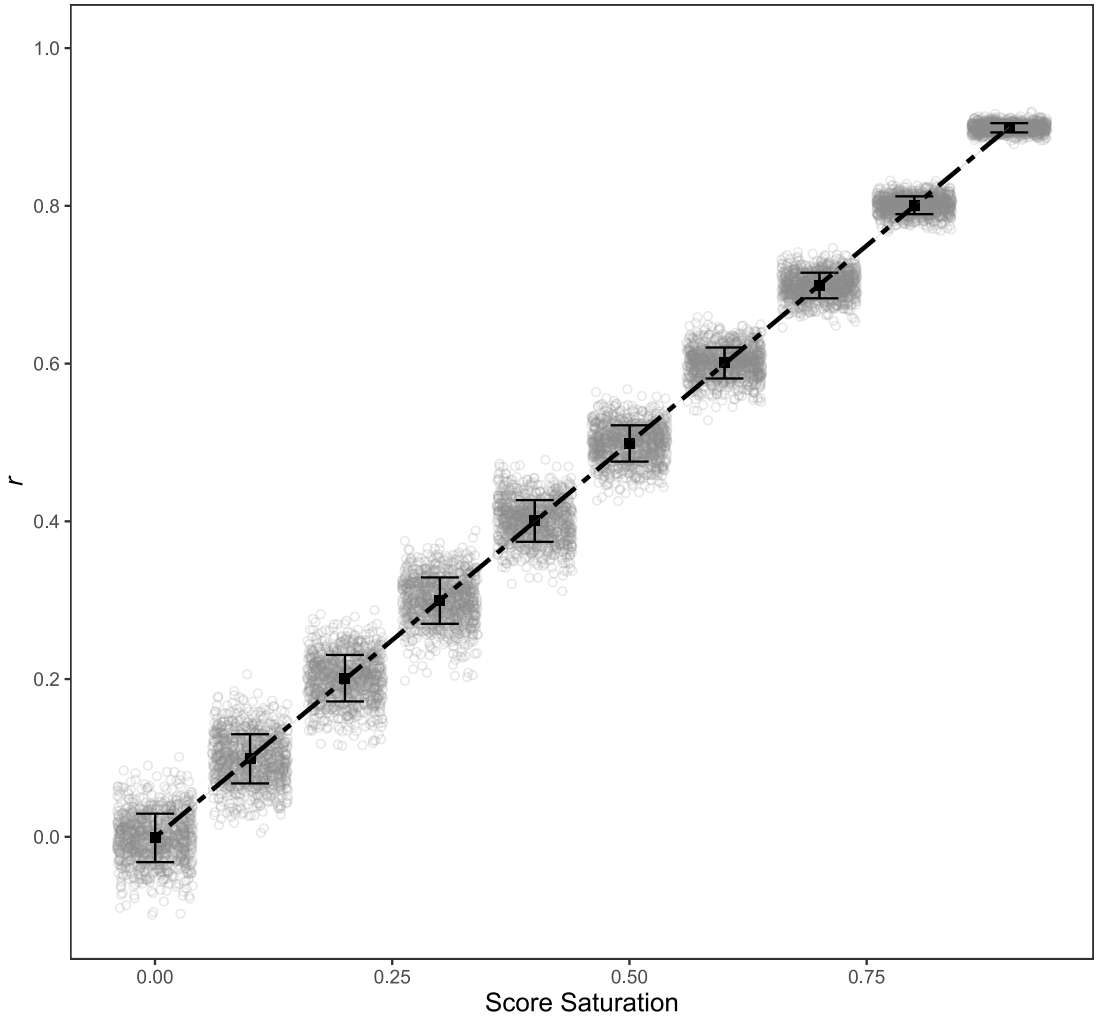
We used the Monte Carlo Technique to simulate response data for SJT items in order to examine sample correlations between SJT and ITP scores. Essentially, SJT scores reflect the congruence of test-takers effectiveness ratings with SMEs' effectiveness ratings, whereas ITP scores reflect the congruence of test-takers effectiveness ratings with SMEs' trait-level ratings. Whether truly effective behavior for a specific SJT has the tendency to reflect a specific trait, can be expressed as the correlation of the effectiveness scoring key and the trait-level scoring key (Motowidlo et al., 2018). We refer to this correlation as score saturation. We automatically generated scoring keys for various, hypothetical SJTs. These SJTs varied in number of items (n_{items} = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100) and score saturation (r_{Scores} = 0, .10, .20, .30, .40, .50, .60, .70, .80, .90). We generated response data with varying sample size (n_{sample} = 50 - 1,000 in steps of 50). Data was simulated under two conditions: multivariate normal distributed data for a latent SJT performance factor and multivariate normal distributed data for a latent ITP factor (αs = 0, .10, .20, .30, .40, .50, .60, .70, .80, .90, 1) with equal means of 0 and standard deviations of 1. For each simulated sample, SJT and ITP scores were computed as within-person correlations between individual response data and the respective scoring key (see Lievens, 2017a). To test the distribution of correlations between SJT and ITP scores, we simulated 1,000 replications for each score saturation. To reduce computing time, however, we did not simulate the full parameter grid. Data for different sample sizes was only simulated with an internal consistency of α = 0 and 50 items. Further, data for different number of items was only simulated for an internal consistency of α = 0 and a sample size of n = 1,000. Finally, data for different internal consistencies of the scores was simulated for SJTs with 50 items and a sample size of n = 1,000. We utilized the R package MASS (version 7.3-51.4) for data generation and the R package MonteCarlo (version 1.0.6) to specify parameter grid loops.

Results

Overall, the observed correlation between SJT and ITP scores converged on average to the scoring key saturation. The average difference between scoring key saturation and observed correlation was $M_{\Delta z}$ = -.00 (SD = .06; see Figure 1). With increasing sample size, the deviation of the observed correlation from the scoring key saturation decreased and eventually stabilized within a corridor of approximately ±.10 (see Figure 2). The number of items had no influence on the correlation between SJT and ITP scores (see Figure 2). All deviations of correlations from

Figure 1

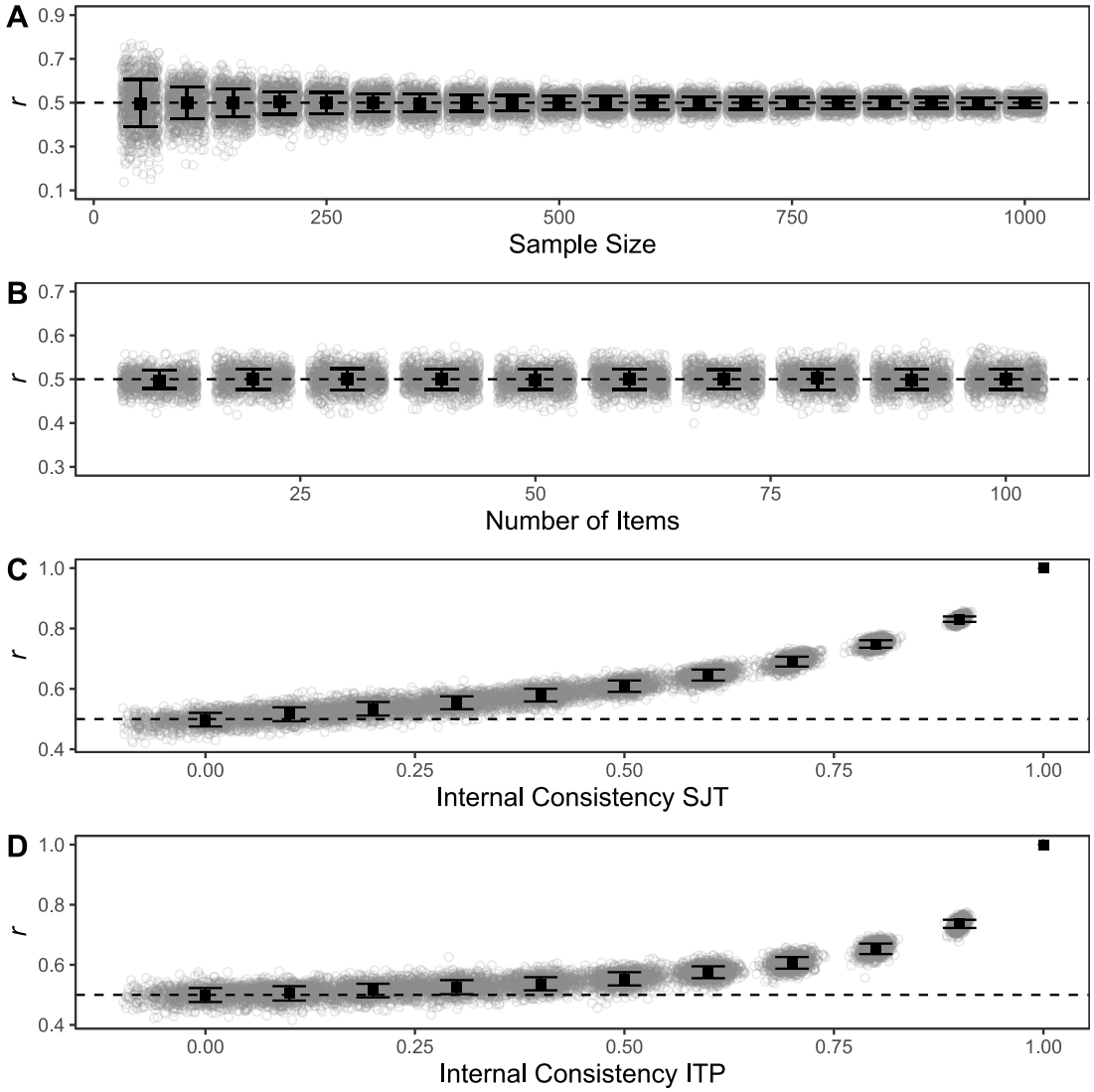
Mean Correlations Between Effectiveness and ITP Scorings (Study 1)



Notes. Plot of correlations between effectiveness and ITP scorings across score saturations for all simulation iterations (N = 1000 and at total of 50 response options). For ach score saturation, mean and standard deviation are depicted. Score saturations were fixed across iterations, thus variation around specific saturation values was included for better visibility of distributions around the mean.

Figure 2

Plots of Results for Varying Simulation Conditions (Study 1)



Notes. Plots depict correlation of effectiveness and ITP scorings against different simulation conditions. **A** Different sample sizes with 50 items and zero internal consistency of test scores. Sample size was fixed to values between 50 - 1,000 in steps of 50 across replications. Variation around those values was added for better visibility of distributions. **B** Different number of items (response options) with $N = 1000$ and zero internal consistency of test scores. Number of items were fixed to values between 10 - 100 in steps of 10 across replications. Variation around those values was added for better visibility of distributions. **C** Different internal consistencies of effectiveness scores (i.e., SJT scores) with $N = 1000$ and 50 items. Empirical internal consistencies are depicted instead of true internal consistencies. **D** Different internal consistency of ITP scores with $N = 1000$ and 50 items. Empirical internal consistencies are depicted instead of true internal consistencies.

the score saturation were attributed to sampling error.

So far, all results referred to data in which all items were essentially unrelated (i.e., $\alpha = 0$). However, even though SJTs often lack internal consistency (e.g., Catano et al., 2012), SJT items show at least some systematic correlation among each other. Our simulation reflected this by successively increasing the internal consistency reliability of the test scores. With increasing reliability of test scores, the correlation between ITP and SJT scores increased and deviated more strongly from the score saturation (see Figure 2). Importantly, this was true irrespective of whether the reliability of the SJT score or the ITP score increased.

Discussion

Study 1 aimed at clarifying to what extent correlations of ITP and SJT scores are confounded when they stem from the same response ratings. This is particularly relevant for SJT research, which stresses the importance of ITPs as underlying process for SJT responses (see Lievens & Motowidlo, 2016; Oostrom et al., 2012). Results of the Monte Carlo simulation showed that correlations of ITP and SJT scores computed from the same responses were tied to the saturation of the two scoring keys. The magnitude of the deviation from the scoring key saturation due to sampling error is in line with research on the general stability of correlation coefficients (Schönbrodt & Perugini, 2013). However, with increasing reliability of any test score, the correlation of ITP and SJT scores increased. In case of unreliable scores, both SJT and ITP scores vary unsystematically around zero. With increasing reliability of any score, more individuals receive higher positive or negative values on the reliable score. Due to the score saturation, both scores will have the same sign (i.e., a positive SJT score will lead to a positive ITP score and vice versa). This systematic match of signs increases the correlation of SJT and ITP scores. The pattern emerges regardless of whether the reliability of the ITP score or the SJT score increases. Hence, no inferences about the true underlying processes are possible. Individuals may utilize either ITPs to respond to SJT items or situation-specific judgements about the true effectiveness of behavioral responses. For instance, Motowidlo and Beier (2010) reported substantial correlations between ITP and SJT scores (.61 - .71), as obtained from the same SJT. However, our results demonstrated that these correlations reflect an artefact due to score saturations (.48 - .57) and score reliabilities (.59 - .65). Therefore, it is necessary to rely on SJTs specifically designed to measure ITPs when examining the relation to other constructs or measures. Next, we scrutinize the construct-related validity of such measures.

Study 2

Sample

An a-priori power analysis ($\alpha = .05$; $1-\beta = .80$) with SemPower (Moshagen & Erdfelder, 2016) for the most complex, possible correlated traits-correlated methods minus 1 model (CTC[M-1]; df = 1636) for measures used in Study 2 revealed a required sample size of $n = 297$ to detect an RMSEA = .05. We expected a dropout rate of 20% and thus collected a total sample of $n = 360$ via the online panel prolific.co. Participants received £4 for completing all tests. We excluded $n = 21$ participants from analyses based on significant Mahalanobis distances (i.e., multivariate outliers; $p < .001$; Meade &

Craig, 2012). We also checked for zero within-person variance and asked participants whether we should use their data for analyses (Meade & Craig, 2012). However, no further participants had to be excluded. Thus, the final data set comprised N = 339 (197 female) cases. Participants were on average M = 39.49 (SD = 10.65; range: 22-65) years old. On average, participants had M = 18.71 (SD = 10.69) years of work experience with M = 36.30 (SD = 9.08) average weekly working hours. For all individuals, at least some of this experience required regular interpersonal interactions (which is important since we used an SJT on teamwork, see below). Most participants were employed in health care (13%), public administration (10%), or retail (9%).

Measures

Situational Judgment Tests

We assessed ITPs for Agreeableness and Conscientiousness with several SJTs. We applied one SJT that was specifically developed to measure ITPs for agreeableness. Additionally, we applied a personality SJT for which the trait expression of each response option was theoretically driven and empirically tested. Finally, we used two SJTs for which the true effectiveness of each response option in the given situation was known. For these two SJTs, SMEs rated the trait expression of all response options. Due to the focus on effectiveness, these SJTs contained response options for which a high (or low) trait expression either reflected effective, ineffective, or neutral behavior. This allowed us to examine whether ITPs reflect general domain knowledge (RQ2). That is, all response options of an SJT may load on a general ITP factor consistently with their trait-level regardless of their effectiveness. This would reflect a trait-specific and context-independent belief. Otherwise, only response options for which trait-level and true effectiveness align should load on an ITP factor. This would reflect ITPs as general domain knowledge.

For all SJT items, we asked participants to rate the effectiveness of response options for the given situation description on a seven-point rating-scale (1 = very ineffective to 7 = very effective). The response instruction and rating scale was adopted from the SJT designed to specifically assess ITPs for agreeableness (Motowidlo et al., 2006b).

Situational Judgment Questionnaire (SJQ). To assess ITPs for agreeableness, we administered an SJT developed by Motowidlo et al. (2006b). The original version of this test consists of 22 situation descriptions about working with people with four response options each. All response options either reflect behavior with high or low trait-level for agreeableness. The SJT was specifically developed by Motowidlo et al. to measure ITPs for agreeableness. To reduce the duration to participate, we applied a six-item short version of this SJT (Freudenstein & Krumm, 2020, see Appendix A). This short version was highly correlated with the long version of the test. Also, both versions correlated virtually identical with self-reported agreeableness. However, the short version had a superior latent model fit. We reverse coded all response options that reflected low agreeableness and then averaged responses within each SJT item. These scores were used as indicators in confirmatory factor analyses (CFA). The unidimensional CFA of this SJT showed an acceptable model fit; $\chi^2(9) = 25.358$, p = .003; CFI = .951; RMSEA = .073; SRMR = .038 (Hu & Bentler, 1999). We computed a manifest test score

by averaging item scores. Reliability of this score was ω = .72.

HEXACO-SJT. We also applied items assessing agreeableness and conscientiousness from the HEXACO-SJT (Oostrom et al., 2018). Each trait in this test is measured with four SJT items that comprise four response options each. Instead of asking participants what they would do in the given situation, we asked them to rate the response options' effectiveness of all eight SJT items. This is the typical procedure to transform an SJT from a measure of personality to a measure of ITPs (see Lievens, 2017a; Motowidlo et al., 2006b). The initial development of the HEXACO-SJT was based on a construct-driven approach, which means that all response options of a given situation description lie on an unidimensional continuum (Guenole et al., 2017; Lievens, 2017b; Oostrom et al., 2018). For the original test development, SMEs rated the trait expressiveness of every response options on scale from -4 to 4 (Oostrom et al., 2018). To score ITPs, we considered all response options for which the average trait expressiveness, as rated by subject matter experts, exceeded |2| (i.e., high or low trait expressiveness). Thus, ratings on 20 out of 32 response options were included to compute ITP scores for agreeableness and conscientiousness (ITP-A and ITP-C). Similar to the SJQ, we reverse coded ratings of response options with low trait expressiveness and averaged ratings within SJT items. The two-dimensional CFA showed a good model fit; $\chi^2(19) = 24.991$, $p = .161$; CFI = .969; RMSEA = .030; SRMR = .033. We averaged ratings on items reflecting agreeableness or conscientiousness to scores of ITPs for the respective trait. Reliabilities of these scores were $\omega_{ITP\text{-}A} = .58$ and $\omega_{ITP\text{-}C} = .52$.

Team Role Test. The Team Role Test (TRT; Mumford et al., 2008) is an SJT assessing team role knowledge. We applied an adapted version of this test, which comprises 10 situation description with four response options each (Schäpers et al., 2019). This SJT's scoring key is based on the response options' effectiveness in the given situation. To develop an ITP scoring key, we asked six PhD students in the field of personality psychology to rate which Big Five trait is reflected by each response option. They also were asked to indicate the corresponding trait level of each response option on a 7-point rating scale. We excluded one rater who failed to select the correct trait for response options of an additional personality SJT item, which we included as a quality check for the raters. Overall, Cohen's κ for the trait ratings was .34. This indicates high ambiguity in the reflected traits of each response options and was to be expected, as the SJT's development was not based on the Big Five taxonomy. Thus, we only considered response options for which at least two third of the raters agreed. This resulted in a Cohen's κ of .69. ICC(2,k) for the trait level ratings was .98. We selected seven SJT items out of which four response options reflected high or low conscientiousness and five response options reflected high or low agreeableness (average ratings < 2.5 or > 5.5). Score saturations between effectiveness scoring keys and ITP scoring keys were $r = .99$ for response options reflecting conscientiousness and $r = .43$ for response options reflecting agreeableness. To compute ITP scores, we reverse coded ratings of response options with low trait expressiveness. However, the two-dimensional model for ITPs for conscientiousness and agreeableness did not converge as the score for ITP-A exhibited a

very low internal consistency (ω = .28). Thus, for all analyses we only considered the ITP score for conscientiousness of this SJT; $\chi^2(2) = 3.708$; $p = .157$; CFI = .973; RMSEA = .050; SRMR = .022; ω = .46. We also computed an effectiveness score for the seven items of this SJT by reverse coding ineffective response options and averaging all responses. Reliability of this score was ω = .72.

Teamwork SJT. Finally, we applied the English version of the Teamwork SJT (TW-SJT; Freudenstein, Remmert, et al., 2020, see Appendix B; Gatzka & Volmer, 2017). Similar to the TRT, this SJT assesses effective teamwork behavior. To determine ITP scoring keys we applied the same procedure as described for the TRT. We asked seven different PhD students to rate response options of this test of which we had to exclude three based on our manipulation check. Overall, Cohen's κ for trait ratings was .39 and .54 for response options with at least two third agreement among raters. ICC(2,k) for the trait level ratings was .82. We included nine SJT items in this study of which seven response options reflected high or low agreeableness and 12 response options reflected high or low conscientiousness. The latent model with all response options did not fit the data; $\chi^2(151) = 389.697$, $p < .001$; CFI = .756; RMSEA = .068; SRMR = .065. Thus, we used Ant Colony Optimization (Olaru et al., 2015; Schultze, 2017) to develop a well-fitting short version with five response options for each ITP factor; $\chi^2(34) = 39.164$, $p < .249$; CFI = .986; RMSEA = .021; SRMR = .035. Reliabilities for the ITP scores for agreeableness and conscientiousness were ω = .53 and ω = .54, respectively. Finally, we computed an effectiveness score by averaging reverse coded ineffective response options and effective response options. Reliability of this score was ω = .62. Score saturations between effectiveness scoring keys and ITP scoring keys were $r = -.11$ and $r = .90$ respectively for response options reflecting conscientiousness and agreeableness.

Self-reported personality

We assessed self-reported agreeableness and conscientiousness with 10 items each taken from the HEXACO-60 (Ashton & Lee, 2009). Participants responded on a seven-point rating scale (1 = strongly disagree to 7 = strongly agree). Reliabilities for the agreeableness and conscientiousness score were ω = .80 and ω = .80, respectively. Additionally, we administered the same 20 items again with a work-specific frame-of-reference. This was done to match the context of all SJT items (Holtrop et al., 2014; Shaffer & Postlethwaite, 2012). Reliabilities for the work-related agreeableness and conscientiousness score were ω = .70 and ω = .78, respectively.

Data Analyses

To test our hypotheses, we first calculated multitrait-multimethod correlation matrices. Second, we fitted CTC(M-1) models (Eid et al., 2003) with the R package lavaan (version 0.6-5; Rosseel, 2012). In this model, a latent trait reference factor is defined by a specific measurement model (e.g., SJQ). All items of other measurements assessing the same trait, load on the same reference factor. Finally, specific measurement factors are defined for all methods except for the reference factor. Latent covariances are restrained to zero between reference and method factors. Importantly, the latent trait factor reflects the meaning of the reference method. Thus, this model allows for computing the shared amount of variance between different methods and the

specific amount of variance of single methods. We estimated two different sets of models. In the first model, the SJQ was set as reference method for agreeableness ITPs and the TRT was set as reference method for conscientiousness ITPs. For all remaining SJTs, specific method factors were modelled. In the second model self-report personality (or self-report work-related personality) was set as reference method. For all SJT items, specific method factors were modelled (see Figure 3).

Results

Bivariate correlations and descriptive statistics are given in Table 2. The average convergent correlation of ITP scores was $r = .30$ ($r = .22$ for agreeableness ITPs and $r = .37$ for conscientiousness ITPs). The average correlation between ITPs for different traits was $r = .21$ (discriminant correlation). Further, ITP-A scores correlated with an average of $r = .23$ with self-reported agreeableness ($r = .26$ with work-related agreeableness). ITP-C scores correlated with an average of $r = .25$ with self-report conscientiousness ($r = .31$ with work-related agreeableness). Average discriminant correlations between ITP scores and self-reported personality were $r = .12$ for ITP-A and $r = .12$ for ITP-C ($rs = .11$ and .19 for correlations to work-related agreeableness and conscientiousness, respectively). Finally, the average correlation between ITP-A scores and SJT effectiveness scores was $r = .11$ and $r = .25$ for ITP-C scores (we excluded ITP scores from the same SJT for this analysis). Correlation coefficients ranged from -.08 to .54 (see Table 2).

The CTC(M-1) with the SJQ and the TRT as reference factors showed a moderate model fit; $\chi^2(325) = 502.100$, $p < .001$; CFI = .875; RMSEA = .040; SRMR = .057 (see Figure 3). In this model, the latent (discriminant) correlation between ITP-A and ITP-C was $r = .51$. On average, 14% of true score variance was shared among ITP-A measures and 86% of true score variance was method specific (see Table 3). The model could explain an average of 62% of the total variance in responses on ITP-A measures. For ITP-C measures, an average of 41% of true score variance was shared among measures and 59% was method specific. An average of 51% of the total variance in responses on ITP-C measures was explained by the model. Notably, the shared variance in ITP-C responses of the TW SJT and the TRT ITP-C factor exceeded the method specific variance (61%). Due to the moderate fit of this model, we also tested a competing model in which all indicators for ITP-A and ITP-C loaded on a common latent factor. We set the SJQ as reference method for all SJTs, regardless of whether they should assess ITP-A or ITP-C. Additionally, method-specific latent factors were specified for all remaining methods. However, we had to merge the method-specific factors for TRT-C and TW-C to achieve model convergence. A Vuong test for comparisons of non-nested models (Vuong, 1989) showed that this model fitted the data slightly better; $z = 1.792$, $p = .037$; $\chi^2(322) = 467.028$, $p < .001$; CFI = .897; RMSEA = .036; SRMR = .048. On average, 12% of the total true score variance of all measures was shared with the SJQ (12% for ITP-A SJTs and 11% for ITP-C SJTs); 88% of the total true score variance was method specific (88% for ITP-A SJTs and 89% for ITP-C SJTs).

Table 2

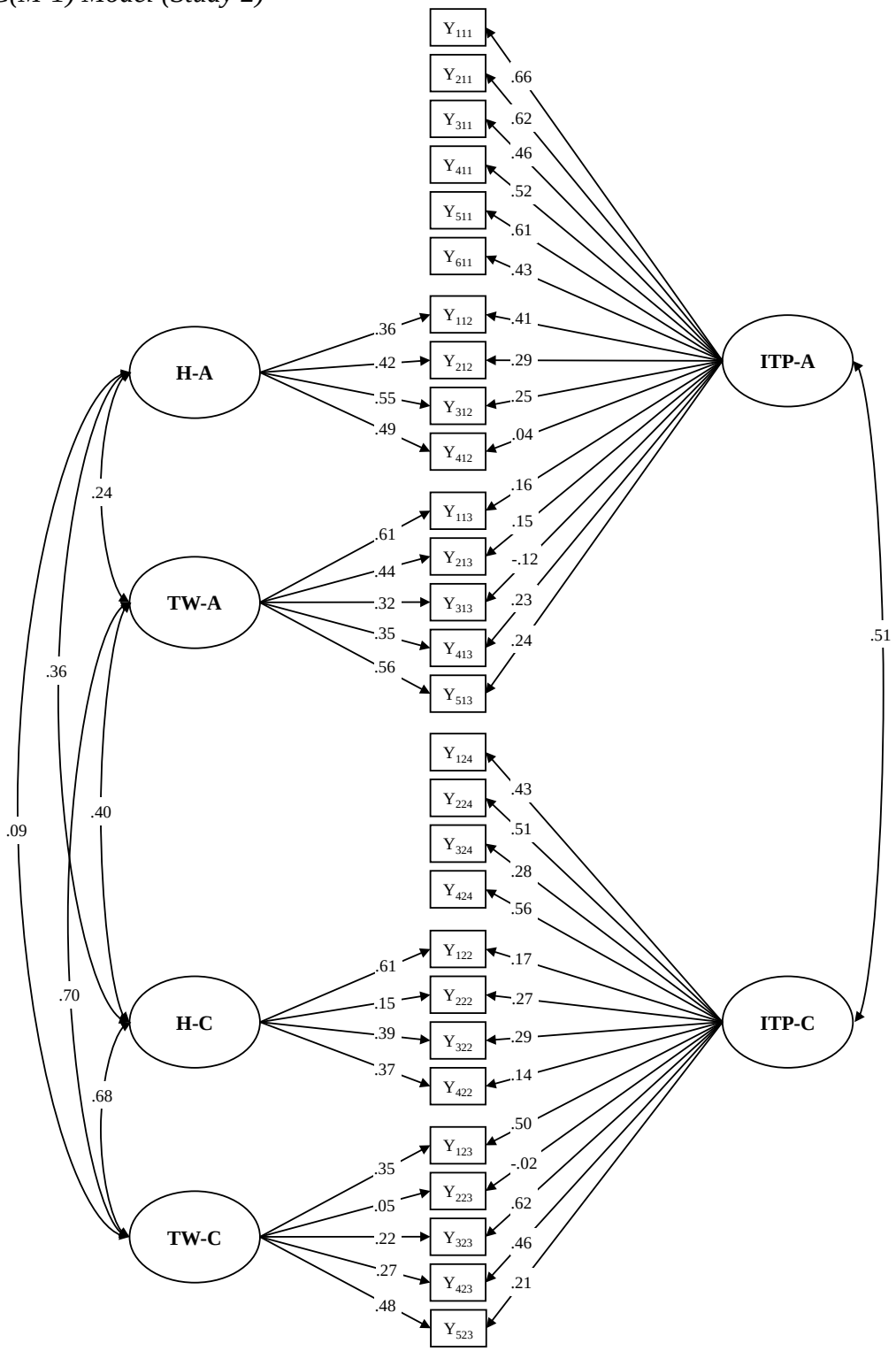
Descriptive Statistics and Bivariate Correlations (Study 2)

	M (*SD*)	1	2	3	4	5	6	7	8	9	10	11
1. SJQ-ITP-A	4.89 (0.56)	-										
2. H-ITP-A	4.13 (0.83)	.33*	-									
3. TW-ITP-A	5.44 (0.72)	.10	.23*	-								
4. TRT-ITP-C	5.44 (0.75)	.25*	.10	.26*	-							
5. H-ITP-C	5.61 (0.70)	.09	.18*	.26*	.33*	-						
6. TW-ITP-C	5.58 (0.65)	.20*	.16*	.38*	.45*	.34*	-					
7. TW effectiveness	4.92 (0.40)	.37*	.00	.14*	.54*	.16*	.36*	-				
8. TRT effectiveness	4.66 (0.57)	.32*	-.05	-.08	.57*	.09	.12*	.61*	-			
9. General A	4.61 (1.01)	.19*	.35*	.13*	.13*	.08	.16*	.06	.04	-		
10. Work-Related A	4.86 (0.83)	.22*	.38*	.16*	.19*	.14*	.25*	.18*	.12*	.81*	-	
11. General C	5.44 (0.88)	.14*	-.04	.17*	.22*	.33*	.19*	.08	.12*	.09	.05	-
12. Work-Related C	5.61 (0.79)	.14*	.00	.20*	.29*	.37*	.27*	.17*	.23*	.08	.12*	.84*

Notes. ITP-A = ITP score for agreeableness; ITP-C = ITP score for conscientiousness; SJQ = Situational Judgment Questionnaire; H = HEXACO SJT; TW = Teamwork SJT; TRT = Team Role Test; A = agreeableness; C = conscientiousness. Please note that correlations between the effectiveness scores and the respective ITP scores are partially confounded (see Study 1). $n = 337 – 338$. * $p < .05$.

Figure 3

CTC(M-1) Model (Study 2)



Notes. CTC(M-1) model with the Situational Judgment Questionnaire as reference method for ITP-A and the Team Role Test as reference method for ITP-C. Thus, ITP-A and ITP-C adopt the meaning of these two methods. $\chi^2(325) = 502.100$, $p < .001$; CFI = .875; RMSEA = .040; SRMR = .057. All coefficients are standardized. Residual variances are not depicted for clarity. ITP = Implicit Trait Policy; A = Agreeableness; C = Conscientiousness; H = HEXACO SJT; TW = Teamwork SJT.

Table 3

Consistent and Specific Variance Components of CTC(M-1) Models (Study 2)

	ITP Consistency		SJT Specificity		General Personality Consistency		Work-Related Personality Consistency	
	Total variance	True variance	Total variance	True variance	Total variance	True variance	Total variance	True Variance
SJQ-ITP-A	72%	100%	-	-	3%	4%	4%	6%
TW-ITP-A	3%	6%	52%	94%	0%	1%	0%	1%
H-ITP-A	13%	22%	46%	78%	13%	23%	12%	21%
TRT-ITP-C	48%	100%	-	-	2%	3%	6%	10%
TW-ITP-C	32%	61%	21%	39%				
H-ITP-C	11%	22%	40%	78%	14%	27%	18%	38%

Notes. ITP-A = ITP score for agreeableness; ITP-C = ITP score for conscientiousness; SJQ = Situational Judgment Questionnaire; H = HEXACO SJT; TW = Teamwork SJT; TRT = Team Role Test; A = agreeableness; C = conscientiousness

Next, we compared all measures of ITPs with self-reported personality. In these models, personality traits were set as reference factors and ITP-specific factors for all SJTs were defined. Hence, we tested how much variance is shared between self-reported personality and ITP measures. Due to a non-positive definite covariance matrix of latent variables, we had to specify a combined method factor for the TRT-C and TW-C items; $\chi^2(555) = 939.498$, $p < .001$; CFI = .847; RMSEA = .045; SRMR = .063 for general personality; $\chi^2(555) = 1002.282$, $p < .001$; CFI = .812; RMSEA = .049; SRMR = .069 for work-related personality. On average, the ITP measures shared 11% of true score variance with the corresponding general self-report personality measures and 15% of true score variance with the corresponding work-related self-report personality measures (see Table 3). Thus, consistent variance components among ITP measures for the same trait were on average higher than consistent variance component of ITP measures and corresponding personality traits. Interestingly, the ITP scores from the HEXACO-SJT shared substantially more variance with the personality measures than all other ITP measures (25% vs. 3% for general personality and 30% vs. 7% for work-related personality). For this SJT, shared variance with self-reported personality was either identical or higher than the shared variance with convergent ITP measures (see Table 3). Thus, we found weak support for H1b overall.

Finally, we took a closer look at the loadings of the TW response options on the ITP reference factors to examine RQ2. If ITPs reflect general domain knowledge, response options that are effective *and* reflect a high trait level as well as response options that are ineffective *and* reflect a low trait-level should have higher loadings on the ITP reference factor than the remaining response options. In fact, the mean loading of effective response options with high trait level expression (and vice versa) was $\lambda = .39$ ($SD = .17$). Response options, for which the true effectiveness was not aligned with the trait expression, had an average loading of $\lambda = .14$ ($SD = .22$) on the ITP factor. We calculated bootstrapped standard errors for the average of unstandardized factor loadings and found that 95% CIs did not overlap [0.71, 1.81; 0.09, 0.50].

Discussion

Study 2 was designed to examine the convergence of different measures for the same ITPs, the divergence of different measures for different ITPs, the relation to self-reported personality, and whether ITPs and reflect general domain knowledge. Our results showed almost no support for Hypotheses 1a and 2. That is, most variance of measures for the same ITPs was not shared across methods but was largely method-specific. However, different SJTs for conscientiousness ITPs showed a larger overlap than SJTs for agreeableness ITPs. Still, for both ITPs, by far the most variance was either due to measurement error or due to method-specific factors. Hence, the operationalization of ITPs with SJTs may not be as straightforward as previously suggested (e.g., Lievens, 2017a; Motowidlo et al., 2006b). This in line with general criticisms of the construct-related validity of SJTs (e.g., McDaniel et al., 2016).

Further, our results showed support for Hypothesis 1b as most measures of ITPs shared more variance with each other than with self-reported personality. Although this generally supports the notion that measures of ITPs are distinct from personality self-reports, these results need to be put into perspective with the generally small amounts of shared variance among ITP measures for the same trait.

Finally, the results showed stronger convergence among response options for which the true effectiveness aligned with the trait level. This speaks to the notion that measures of ITPs reflect general domain knowledge (see Lievens & Motowidlo, 2016). However, large proportions of variance were also method-specific (i.e., SJT-specific). Thus, SJTs were no pure measures of general domain knowledge. Instead, they mostly assessed a test-specific construct. Overall, Study 2 did not provide evidence for the construct-related validity of ITPs as measured with SJT items.

General Discussion

Across two studies, we sought to assess the construct-related validity of ITPs. Thereby, we responded to recent calls for an integration of ITPs into personality research (see Lievens, 2017a). The Monte Carlo simulation conducted in Study 1 showed that variance in SJTs usually attributed to ITPs might be an artefact of ITP saturation (i.e., coinciding trait and effectiveness ratings). From this follows that separate measures are needed to investigate ITPs. In Study 2, we pursued this approach, but demonstrated that measures which were specifically designed to gauge the same ITPs lack

systematic overlap. In other words, we found only weak convergent correlations. Convergent correlations were slightly higher than discriminant correlations. Finally, correlations of ITP scores and personality were smaller when compared to the convergent correlations. Generally, differences among convergent and discriminant correlations were small. In light of the lack of evidence for convergent validity, we question the construct-related validity of measures of ITPs. However, the results also contain fine-grained nuances that need further discussion against the background of previous research.

First, ITPs are conceptualized to be trait-specific (Motowidlo et al., 2006a, 2006b). That is, implicit beliefs about the effectiveness of behavior should be defined by the underlying trait of the behavior. Our results do not support this perspective. Although there was some convergence across measures of the same ITP, this overlap did not or only marginally exceed the variance shared by measures of different ITPs. In fact, a latent model with only a single ITP factor did not show a worse model fit than the competing two-factorial model. Interestingly, these results contradict previous research that assessed ITPs for several traits (Motowidlo et al., 2006b). However, the current research is the first to include a multitrait-multimethod approach to validate ITP measures.

Second, personality has been described as an antecedent of ITPs. Yet, personality and ITPs are considered distinct constructs (Lievens, 2017a; Motowidlo et al., 2006b). In general, our results support this point of view. Although measures of ITPs share some variance with measures of personality, most of the variance in ITPs is distinct from self-reported personality. This is in line with previous studies examining the relation of ITPs and personality (Martin-Raugh et al., 2016; Motowidlo et al., 2006b, 2016, 2018). However, the magnitude of shared variance between ITPs and personality needs to be contrasted against the amount of shared variance among ITPs for the same trait. For instance, the differences of unique true score measures of ITPs for agreeableness and shared variance with self-reported personality was negligible. Measures of ITPs for conscientiousness, however, showed an expected pattern in that more variance was shared within methods of ITPs than with self-reported personality. The ITP measures that were based on the HEXACO-SJT were the exception as they shared a substantial amount of variance with self-reported personality. These results suggest that the relation of ITP measures and personality is SJT specific.

Third and related to the second point, ITPs are conceptualized as context-independent (Lievens & Motowidlo, 2016). That is, ITPs reflect *general* beliefs about the effectiveness of behaviors that are shaped by the underlying trait and not situational influences. However, our results demonstrated that by far the largest amount of true score variance was attributed for by specific methods (i.e., SJTs). In fact, these results are very similar to previously reported bivariate correlations between ITP scores for several SJTs (Motowidlo et al., 2018). Thus, an individual's tendency to rate specific behaviors as effective was mostly driven by the specific method and therefore specific contextual influences. This is in line with recent research demonstrating that situational processes take place when responding to SJT items, even if situation descriptions were omitted (Freudenstein, Schäpers, et al., 2020).

Fourth, we found preliminary support for the notion of ITPs as general domain knowledge. That is, indicators of ITPs correlated more strongly, when the trait expression was in line with the true effectiveness of these behaviors. In other words, test-takers did not utilize general beliefs about trait-related behaviors irrespective of the true effectiveness to respond to SJT items. However, the SJTs applied in this study were also no valid measures of trait-related general domain knowledge. This is reflected in the large variance components that were specific to the respective SJTs. Overall, these results seem insufficient to draw conclusions about the conceptualization of ITPs. Theoretical arguments are needed to clearly define ITPs as general beliefs or as general domain knowledge.

Importantly, the construct of ITPs was originally developed to explain why SJTs correlate with personality traits (Motowidlo et al., 2006b), which is why the assessment of ITPs is exclusively bound to SJTs. However, our results show that this assessment approach to ITPs lacks construct-related validity. This lack of construct-related validity questions the relevance of ITPs for SJT theory and beyond (cf. Lievens & Motowidlo, 2016). For instance, against the background of our findings, previous links between ITP scores and role-play behaviors (Martin-Raugh et al., 2016; Motowidlo et al., 2006b) may rather be evidence for the criterion-related validity of the specific SJTs than the predictive validity of ITPs.

Notably, our results are in line with Pretsch and Schmitt (2017), who argued that typical SJTs are generally not suitable for the assessment of ITPs. Specifically, they pointed out that current assessments of ITPs lack a systematic combination of behaviors that reflect several traits and several possible consequences. Such a systematic combination may indeed be helpful to revive research on ITPs by rethinking assessment methods. In any case, sound ITP assessments have to be established before the outline by Lievens (2017a) on research questions for adopting ITPs into personality research may be pursued.

Another necessary step to facilitate research on ITPs is to extend our knowledge about relations to associated constructs. For instance, external beliefs or other implicit measures may be helpful to understand ITPs (Lievens, 2017a). More thorough investigation of the nomological net of ITPs may also clarify conceptual caveats such as the distinction of ITPs as beliefs and general domain knowledge.

Limitations

We would like to stress that our results do not suffice to draw conclusions about the general theory of ITPs. That is, the above arguments only apply to the current way of measuring ITPs. However, any conclusions about the existence or usefulness of the concept of ITPs requires a method that allows valid inferences about the underlying construct. The SJTs applied in our study (except for one) were originally developed to serve another purpose than the assessment of ITPs. Although every SJT has been deemed suitable for the assessment of ITPs, specific measures exist that were designed to assess ITPs (e.g., single response SJTs; Motowidlo et al., 2016). Future research should investigate the usefulness of these measures regarding the assessment of ITPs.

Conclusions about the general theory of ITPs and beyond the assessment of ITPs are further limited due to the

restricted nomological net of ITPs included in our study. Specifically, we only assessed self-reported personality. As delineated above, other constructs may also be of interest when studying ITPs. However, examining relations of ITPs and similar constructs has not been done beyond personality traits. In fact, no other relevant construct besides personality is explicitly incorporated into ITP theory. Future research should incorporate ITPs into similar frameworks about the relation of basic needs, beliefs, personality and behavior (e.g., Dweck, 2017).

Conclusion

In response to recent calls to incorporate ITPs into personality research, the current research sought to examine whether meaningful inference can be drawn from measures of ITPs. First and foremost, our results demonstrated that correlations between ITPs and SJTs derived from the same responses depend on ITP saturation and thus are a statistical artefact. Moreover, we showed that the majority of ITP measures lacked construct-related validity, thereby questioning their usefulness to explain SJT scores. That is, most of the observed variance was either method specific or due to measurement error. Hence, more research is needed that examines whether ITPs can be assessed validly before conclusions can be drawn about the benefits of this construct for the field of personality psychology.

References

Ashton, M., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, J. J. A., Fleeson, W., Grafton, B., Jayawickreme, E., Kurzius, E., MacLeod, C., Miller, L. C., Read, S. J., Roberts, B., Robinson, M. D., Wood, D., Wrzus, C., & Mõttus, R. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality*, *31*(5), 503–528. https://doi.org/10.1002/per.2115

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*(3), 223–235. https://doi.org/10.1111/j.1468-2389.2006.00345.x

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

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*(3), 333–346. https://doi.org/10.1111/j.1468-2389.2012.00604.x

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*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Dweck, C. S. (2017). From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, *124*(6), 689–719. https://doi.org/10.1037/rev0000082

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods*, *8*(1), 38–60. https://doi.org/10.1037/1082-989x.8.1.38

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, *75*(4), 825–862.

https://doi.org/10.1111/j.1467-6494.2007.00458.x

Freudenstein, J.-P., & Krumm, S. (2020). *Developing a short-form situational judgment test to assess implicit trait policies for agreeableness*. https://doi.org/10.31219/osf.io/kax7n

Freudenstein, J.-P., Remmert, N., Reznik, N., & Krumm, S. (2020). *English translation of the teamwork situational judgment test (SJT-TW)* [Manuscript submitted for publication].

Freudenstein, J.-P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. https://doi.org/10.1111/peps.12385

Funder, D. C. (2016). Taking situations seriously: The situation construal model and the Riverside Situational Q-Sort. *Current Directions in Psychological Science*, *25*(3), 203–208. https://doi.org/10.1177/0963721416635552

Gatzka, T., & Volmer, J. (2017). Situational Judgment Test für Teamarbeit (SJT-TA) [situational judgment test for teamwork]. In *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis249

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.

Guenole, N., Chernyshenko, O. S., & Weekly, J. A. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, *17*(3), 234–252. https://doi.org/10.1080/15305058.2017.1297817

Holtrop, D., Born, M. P., de Vries, A., & de Vries, R. E. (2014). A matter of context: A comparison of two types of contextualized personality measures. *Personality and Individual Differences*, *68*, 234–240. https://doi.org/10.1016/j.paid.2014.04.029

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

Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2016). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology*, *90*(1), 1–27. https://doi.org/10.1111/joop.12151

Judge, T. A., Hofmans, J., & Wille, B. (2017). Situational judgement tests and personality measurement: Some answers and more questions. *European Journal of Personality*, *31*(5), 463–464. https://doi.org/10.1002/per.2119

Kronmal, R. A. (1993). Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *156*(3), 379–392. https://doi.org/10.2307/2983064

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*(2), 399–417. https://doi.org/10.1037/a0037674

Lievens, F. (2017a). Assessing personality-situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*(5), 424–440. https://doi.org/10.1002/per.2111

Lievens, F. (2017b). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing*, *17*(3), 269–276. https://doi.org/10.1080/15305058.2017.1309857

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

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*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Martin-Raugh, M. P., Kell, H. J., & Motowidlo, S. J. (2016). Prosocial knowledge mediates effects of agreeableness and emotional intelligence on prosocial behavior. *Personality and Individual Differences*, *90*, 41–49. https://doi.org/10.1016/j.paid.2015.10.024

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: Perspectives on Science and Practice*, *9*(1), 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*(4), 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*(1-2), 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*(3), 437–455. https://doi.org/10.1037/a0028085

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. https://doi.org/1995-25136-001

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 54–60. https://doi.org/10.1080/10705511.2014.950896

Motowidlo, S. J. (2003). Job Performance. In W. C. Borman & R. J. Klimoski (Eds.), *Handbook of Psychology: Industrial and Organizational Psychology, Vol. 12* (pp. 39–54). John Wiley & Sons Inc.

Motowidlo, S. J. (2017). Implicit Trait Policies in Personality Research. *European Journal of Personality*, *31*(5), 472–473. https://doi.org/10.1002/per.2119

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*(2), 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*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance*, *29*(4), 331–346. https://doi.org/10.1080/08959285.2016.1165227

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). 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). Lawrence Erlbaum Associates.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*(4), 749–761. https://doi.org/10.1037/0021-9010.91.4.749

Motowidlo, S. J., Lievens, F., & Ghosh, K. (2018). Prosocial implicit trait policies underlie performance on different situational judgment tests with interpersonal content. *Human Performance*, *31*(4), 238–254. https://doi.org/10.1080/08959285.2018.1523909

Motowidlo, S. J., & Peterson, N. G. (2008). Effects of organizational perspective on implicit trait policies about correctional officers' job performance. *Human Performance*, *21*(4), 396–413. https://doi.org/10.1080/08959280802347197

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*(2), 250–267. https://doi.org/10.1037/0021-9010.93.2.250

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*(4), 335–353. https://doi.org/10.1080/08959285.2012.703732

Oostrom, J. K., de Vries, R. E., & de Wit, M. (2018). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*(1), 1–29. https://doi.org/10.1080/08959285.2018.1539856

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, *60*(359–367), 489–498.

https://doi.org/10.1098/rspl.1896.0076

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), 1–16. https://doi.org/10.1111/1468-2389.00222

Pretsch, J., & Schmitt, M. (2017). Validity risks and potential advancements of situational judgment tests and assessment centre exercises in personality research. *European Journal of Personality*, *31*(5), 477–478. https://doi.org/10.1002/per.2119

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677–718. https://doi.org/10.1037/a0037250

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schäpers, P., Freudenstein, J.-P., Mussel, P., Lievens, F., & Krumm, S. (2020). Effects of situation descriptions on the validity of construct-driven situational judgment tests. *Journal of Research in Personality*.

Schäpers, P., Lievens, F., Freudenstein, J.-P., Hüffmeier, J., König, C. J., & Krumm, S. (2020). 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*, *93*(2), 472–494. https://doi.org/10.1111/joop.12297

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

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schultze, M. (2017). *Constructing subtests using ant colony optimization* [Doctoral dissertation, Freie Universität Berlin]. http://doi.org/10.17169/refubium-622

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*(3), 445–494. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307–333. https://doi.org/10.2307/1912557

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 295–322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

Wright, A. G. (2017). Leveraging situational judgement tests to study pathological personality processes. *European Journal of Personality*, *31*(5), 485–486. https://doi.org/10.1002/per.2119

Chapter 3

Is It All in the Eye of the Beholder? The Importance of Situation Construal for Situational Judgment Test Performance

This article was previously published as:

Freudenstein, J.-P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. https://doi.org/10.1111/peps.12385

Is It All in the Eye of the Beholder? The Importance of Situation Construal for Situational Judgment Test Performance

Jan-Philipp Freudenstein
Freie Universität Berlin

Philipp Schäpers
Singapore Management University

Lena Roemer
Humboldt-Universität zu Berlin

Patrick Mussel & Stefan Krumm
Freie Universität Berlin

Recent research challenges the importance of situation descriptions for Situational Judgment Test (SJT) performance. This study contributes to resolving the ongoing debate on whether or not 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 re-conceptualize 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. Situational Judgment Test, Situation Construal, Person × Situation Interaction, Validity

Correspondence concerning this paper should be addressed to Jan-Philipp Freudenstein, Institute of Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany. E-Mail: jan-philipp.freudenstein@fu-berlin.de

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). The supporting information is available online (doi.org/10.1111/peps.12385) and in Appendix D.

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 and colleagues 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. 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

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

(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.

Theoretical Background

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., 2015; Weekley et al., 2006; 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. 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 and colleagues 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. 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 non-agreeable 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-*in*dependent measures (i.e., measures of general domain knowledge).

Evidence in Favor of the Situation

Despite the aforementioned evidence and recent calls for a re-conceptualization 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. 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. 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 inter-individual 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).

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, Sherman, & Funder, 2015). They represent an individual's psychological construal of the situation and encompass the interaction process between situational cues and inter-individual 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, Sherman, & Funder, 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

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 were 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).

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, Sherman, Nave, 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 inter-individual 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, Sherman, & Funder, 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, i.e., 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; Leeds, 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) as 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 re-construe 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. 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., 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 situation 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 et al., 2017; Rauthmann et al., 2014; Sherman et al., 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.

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 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).

Study 1

Methods

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 female, 4 other) participants were included in subsequent statistical analyses. On average, participants' age was M = 24.58 years (SD = 5.52, range 18 to 66). Almost all participants held a university

[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).

entrance diploma (95%). Furthermore, 33% of the sample had at least an undergraduate degree and additional 12% had completed vocational education and training (VET; three years of vocational training and education for skilled crafts and trades within the German dual system).

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.

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, i.e., 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 $\alpha = .57$ and $\omega = .57$[4,5]. A sample item can be found in the online supplementary material.

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 supplementary material 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 $\alpha = .67$ and $\omega = .70$. A sample item can be found in the online supplementary material.

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 $\alpha = .31$ and $\omega = .34$. A sample item can be found in the online supplementary material.

Perceived situation characteristics. Rauthmann et al. (2014) developed a taxonomy of perceived situation

[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.

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 et al., 2014; Rauthmann & Sherman, 2016a) and exhibits substantial predictive validity for individual behavior over and above personality (Parrigon, Woo, Tay, & Wang, 2017; Rauthmann et al., 2014; Rauthmann & Sherman, 2016a, 2016b; Rauthmann, Sherman, Nave, et al., 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.

Data Analyses. Since 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,

Table 1

Pooled descriptive statistics of the DIAMONDS across SJT items

	***M* (*SD*)**	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**
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
10. Group 2	-	.08	.04	.18	.05	.18	-.07	.04
11. Group 3	-	-.06	-.01	-.07	-.01	-.16	.14	.04
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

Notes. Sample sizes in Study 1 ranged between $n = 209 – 224$. Sample sizes in Study 2 ranged between $n = 561 – 632$, $n_{\text{group 1}} = 261$; $n_{\text{group 2}} = 214$; $n_{\text{group 3}} = 252$. Sample sizes in Study 3 ranged between $n = 284 – 285$.

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, Rauthmann, Brown, Serfass, & Jones, 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 in the online supplementary material.

Results

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.

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, $\Delta\chi^2(1) = 1566.10$, $p < .001$, $R^2_{McF/H} = .143$. The same was true for the random intercept for subjects, but only if adjusted for the SJT items' nested structure within three different SJTs, $\Delta\chi^2(6) = 554.19$, $p < .001$, $R^2_{McF/H} = .050$. 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 out 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, $\Delta\chi^2(52) = 890.32$, $p < .001$, $R^2_{McF/H} = .090$. In sum, these results lend support to Hypothesis 1.

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).

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.

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	.07	.06	.290	1.07	.07*	<.001	.05	.459	-.08							
I	.09	.06	.126	1.09	.05*	<.001	.07	.261	.18	-.16						
A	.15*	.05	.002	1.16	.03*	<.001	.08	.204	.24	-.56	-.26					
M	-.05	.05	.303	0.95	.01	.663	-.12*	.030	-.11	.52	-.62	-.05				
O	-.32*	.09	.001	0.72	.17*	<.001	-.22*	.003	-.38	-.04	.37	-.70	-.51			
N	.14*	.07	.040	1.15	.08*	<.001	.09	.085	.20	-.34	-.34	.66	.37	-.81		
De	-.11*	.05	.011	0.89	.02	.435	-.06	.397	.15	.30	.63	-.81	-.40	.75	-.79	
S	.11	.06	.052	1.11	.05*	<.001	.05	.191	-.21	-.07	.01	-.27	.45	.18	-.07	.10
logLikelihood -4189.89																

Notes. 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. D = Duty, I = Intellect, A = Adversity, M = Mating, O = pOsitivity, N = Negativity, De = Deception, S = Sociality. $N = 227$.
* $p < .05$, *p*-values for random effects refer to likelihood-ratio tests between models with and without the corresponding parameter.

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 (since 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).

Methods

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 non-significant 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 female, 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. Whereas 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%).

Study Design and Materials. All data were collected online in a between-subjects design with participants randomly assigned to three groups ($n_{\text{group 1}}$ = 261; $n_{\text{group 2}}$ = 214; $n_{\text{group 3}}$ = 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 vs. without situation descriptions (i.e., description-dependent items; for examples, see online supplementary material). 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

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

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 supplementary material.

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 supplementary material.

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 $\alpha = .41$ to $\alpha = .71$.

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 multi-group 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 in the online supplementary material.

Results

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 in the online supplementary material). Therefore, we removed these three items from subsequent analyses. Notably, the interpretation of the main results presented below did not differ when re-categorizing the three items (see Table S5 in the online supplementary material).

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. Since 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 in the online supplementary information).

Hypothesis Tests. Overall, in all three groups, at least one dimension of DIAMONDS significantly predicted performance for eight out 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* = .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 a 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.

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 or not 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.

Table 3

Multi-group 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*	G1	G2	G3	G1	G2	G3
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	group 1 & 2	5.767	3.92	.209	I, De, S	I, De, S	S	.25	.22	.14
PI 10	55.987	<.001	group 1 & 2	8.762	3.60	.052	D, I, S	D, I, S	O, N	.18	.20	.08
FB 8	51.704	<.001	group 1 & 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

Notes. 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. TRT = Team Role Test, PI = personal initiative SJT, FB = Facebook SJT, D = Duty, I = Intellect, A = Adversity, M = Mating, O = pOsitivity, N = Negativity, De = Deception, S = Sociality, G = Group. Sample sizes ranged between $n_{group1} = 205 – 234$, $n_{group2} = 142 – 170$, $n_{group3} = 211 – 230$. * $p < .05$

However, no substantial correlations were found ($r_{group\ 1}$ = -.007, p = .983; $r_{group\ 2}$ = -.004, p = .991; $r_{group\ 1}$ = -.025, p = .940).

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 inter-rater reliability was *ICC*2 = .71 for the work-related items (Figure S7 in the online supplementary material 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 - .37), r_{group2} = .10 (SD_{group2} = .09; range: -.19 - .58), r_{group3} = .09 (SD_{group3} = .08; range: -.20 - .39). A two-way ANOVA (group × description dependency of SJT items) revealed that the profile correlations differed across groups, $F(2, 4655) = 26.33$, $p < .001$, η^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 in the online supplementary material) 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 out of seven SJT items

[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 *ICC*2 = .27. Thus, we only inspected ratings for the work-related SJT items.

(B_{TRT5} = -0.27, 95% CI [-0.43, -0.14], β_{TRT5} = -.09; B_{PI9} = -0.12, 95% CI [-0.22, -0.02], β_{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 out 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 supplementary material, 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.

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 out 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 situations 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 the experts. This is in line with Rauthmann et al.'s (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).

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.

Methods

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 inter-individual differences. After exclusion of careless responders, $N = 285$ (174 female, 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).

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. Afterwards, 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 α = .65 and ω = .66. The administration of the SJT measuring self-consciousness was identical to Study 1. The reliability of this SJT was α = .69 and ω = .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 Study 1 and Study 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 α = .50 to α = .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 α = .82 for personal initiative and α = .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

[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.

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 α_{IRB} = .89, α_{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 .30 to .67. We also assessed self-rated IRB and OCBI (α_{IRB} = .81, α_{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 α = .78 and α = .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 α = .82 to α = .90. We computed a score for general mental ability following the test authors' instructions. Reliability of this score was α = .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 α = .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 α = .67 to α = .81.

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 in the online supplementary material.

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 (.61)	.18*	-.01	*(.82)*												
4. SC Peer	2.86 (.93)	-.14	.02	-.26*	*(.76)*											
5. OCBI Peer	4.16 (.64)	-.07	-.01	.58*	-.05	*(.87)*										
6. IRB Peer	4.49 (.60)	.01	.03	.58*	-.07	.76*	*(.89)*									
Other constructs																
7. PI Self	3.65 (.61)	.36*	-.22*	.26*	-.12	.05	.05	*(.78)*								
8. SC Self	3.50 (.98)	-.24*	.50*	-.05	.25*	.06	.07	-.37*	*(.70)*							
9. ES	3.12 (.87)	.17*	-.26*	.06	-.28*	-.07	-.04	.32*	-.60*	*(.75)*						
10. E	3.56 (.87)	.30*	-.32*	.13	-.16	.05	-.01	.38*	-.48*	.28*	*(.81)*					
11. C	3.75 (.71)	.24*	-.07	.32*	-.08	.04	.14	.47*	-.19*	.19*	.21*	*(.69)*				
12. A	3.11 (.84)	.08	-.08	.04	.00	.13	.10	.08	-.23*	.21*	.22*	.02	*(.67)*			
13. O	3.99 (.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)*

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

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	-	.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

Notes. 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. PI = personal initiative, SC = self-consciousness, OCBI = organizational citizenship behavior, IRB = in-role behavior, D = Duty, I = Intellect, A = Adversity, M = Mating, O = pOsitivity, N = Negativity, De = Deception, S = Sociality. *n =125*.

Results

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 out 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 ≈ .0028; Cabin & Mitchell, 2000), the link between perceived situation characteristics and SJT responses remained significant for six SJT items. In a 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,

[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.

DIAMONDS explained ΔR^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 ΔR^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 out of 18 SJT items (seven items when corrected for alpha inflation).

Hypothesis Tests. Overall personal initiative SJT scores predicted peer-rated personal initiative (β = .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 out of 18 SJT items, perceived situation characteristics significantly predicted at least one peer-rated criterion above and beyond SJT item performance, with average ΔR^2`s 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 in the online supplementary material). Generally, a similar picture emerged for self-rated criteria (Table S10 in the online supplementary material). 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], β_{N2} = .10) and Negativity (B_{N3} = 0.07, 95% CI [0.01, 0.14], β_{N3} = .07) mediated the relation between self-reported self-consciousness and SJT item responses.

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 non-linear interaction processes between person and situation may be fruitful for further investigations (e.g., Blum et al., 2018).

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.

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 non-situational. 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., 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.

Implications for Practice and Research

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. Since 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 et al., 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).

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.

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.

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 III, Mueller-Hanson, & Rupp, 2017), additional information about test-takers' construal may help uncover the reasons for ineffective behavior.

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 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.

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 re-conceptualize 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.

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. 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, & 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/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. *The 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 GmbH.

Chapter 4

The Influence of Situational Strength on the Relation of Personality and Situational Judgment Test Performance

This article has been prepared for publication:

Freudenstein, J.-P., Schäpers, P., Reznik, N., & Krumm, S. (2020). *The influence of situational strength on the relation of personality and situational judgment test performance* [Manuscript prepared for publication].

The Influence of Situational Strength on the Relation of Personality and Situational Judgment Test Performance

Jan-Philipp Freudenstein
Freie Universität Berlin

Philipp Schäpers
Singapore Management University

Nomi Reznik & Stefan Krumm
Freie Universität Berlin

Situational strength theory has been used as theoretical underpinning of person-situation processes that are linked to job performance. Accordingly, the link between personality traits and job performance increases in weak situations. Building on this research, similar mechanisms have been proposed for simulation-based selection tools, such as Situational Judgment Tests (SJTs), to explain how these measures work as predictors of job performance. However, underlying processes of SJT performance are subject to debate with some scholars arguing in favor of context-independent processes while others maintain that situations play an essential role. This study (N = 707) examined whether the strength of situations in SJT items moderated the relation of personality and SJT performance. Results did not support the notion that personality is more strongly related to SJT performance when situations are weak. In fact, for some traits the opposite may be true as more situational constraints led to an increase in the relation of extraversion, emotional stability, and SJT performance. The results add to an increasing body of research about psychological processes in SJTs. Limitations and implications for research and practice are discussed.

Keywords: Situational Strength, Situational Judgment Tests, Personality

In line with general assumptions about underlying processes of individual behavior, person–situation interactions contributed to an increased understanding about predictions of job performance (e.g., Judge & Zapata, 2015; Meyer et al., 2009; Tett & Burnett, 2003). In particular, situational strength theory offered insights about the relation of personality and behavior at work: Strong situations will lead to more homogeneous behavior across individuals resulting in weaker links between personality and job

Correspondence concerning this paper should be addressed to Jan-Philipp Freudenstein, Institute of Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany. E-Mail: jan-philipp.freudenstein@fu-berlin.de

The German Research Foundation (KR 3457/2-3) funded this research.

Data and R code are available on the Open Science Framework (osf.io/hck3j).

performance (e.g., Judge & Zapata, 2015; Meyer et al., 2009). Accordingly, similar underlying processes have been incorporated into the development of personnel selection tools such as assessment centers (e.g., Lievens et al., 2009; Melchers et al., 2012; Schollaert & Lievens, 2012).

Similarly, Situational Judgment Tests (SJTs) are popular selection tools that rest on the principle of person-situation interaction and behavioral consistency (Lievens & De Soete, 2012; Motowidlo et al., 1990; Weekley et al., 2015). That is, SJTs typically comprise various work-related situation descriptions and several short behavioral response options (McDaniel & Nguyen, 2001). Despite recent efforts to include an interactionist perspective into SJT theory (e.g., Campion & Ployhart, 2013; Harris et al., 2016; Martin-Raugh & Kell, 2019), the underlying processes of SJT performance remain subject to debate. This is due to recent studies demonstrating that situation descriptions in SJTs are often not relevant for SJT performance (Krumm et al., 2015; Schäpers et al., 2019; Schäpers, Freudenstein, et al., 2020; Schäpers, Lievens, et al., 2020). These authors argued that SJTs are less dependent on person-situation processes than previously assumed.

On the other hand, a recent study upheld person-situation interaction as underlying processe of SJTs (Freudenstein, Schäpers, et al., 2020). These authors demonstrated that an individual's perception of situations in SJT items is relevant for responses to SJT items. The results further showed that situation descriptions and response options jointly constitute psychologically relevant situations and that stripping of situation descriptions does not transform SJTs into context-independent measures (see also Harris et al., 2016; Melchers & Kleinmann, 2016). In the current research, we further shed light onto the person-situation interplay that underlies SJT performance. Specifically, we examine whether the situational strength of SJT items moderates the relation of personality and SJT performance (see Harris et al., 2016). By doing so, we contribute to an understanding of SJTs' functioning and why SJTs predict relevant criteria.

Theoretical Background

One inference from the person-situation debate was that behavior is driven by both persons and situations (e.g, Fleeson & Noftle, 2008). Situational strength forms one type of such situational influences on behavior (e.g., Dalal et al., 2014; Judge & Zapata, 2015; Meyer et al., 2009, 2010; Mischel, 1973, 1977). Situational strength is defined "as implicit or explicit cues provided by external entities regarding the desirability of potential behaviors" (Meyer et al., 2010, p. 122). Accordingly, strong situations should lead to more similar perceptions of situations and thus more similar behavior (Meyer et al., 2009; Mischel, 1977). To clarify the conceptual framework of situational strength, Meyer et al. (2010) proposed four facets, namely clarity, consistency, constraints, and consequences. Each facet describes a group of situational cues that restrict the range of possible behavior (i.e., clarity of responsibilities, consistency of different situational demands, constraints to behavior, and consequences of behavior). For instance, "clear instructions and support from one's supervisor" should increase the awareness about expected behavior for all employees (Meyer et al., 2010, p. 125).

It is assumed that with increasing

situational strength, the situation becomes more relevant as a determinant of behavior, which is accompanied by a decrease in the relevance of personality as predictor of behavior (Meyer et al., 2009). Several studies supported this role of situational strength as negative moderator on the relation of personality traits and job performance (Judge & Zapata, 2015; Lee & Dalal, 2016; Meyer et al., 2009). Meyer et al. demonstrated in their meta-analysis that conscientiousness correlated more strongly with performance criteria in weak occupations. Accordingly, research adopted the concept of situational strength to simulation-based tools in personnel selection (e.g., assessment center; Christiansen et al., 2013; Herde & Lievens, 2018; Lievens et al., 2009, 2015; Melchers et al., 2012; Oliver et al., 2016; Schollaert & Lievens, 2012). From this it follows that simulation-based selection measures should refrain from incorporating strong situations to maximize the assessment of relevant traits or dimensions that predict job performance (e.g., Lievens et al., 2009; Melchers et al., 2012).

In line with this proposition, Harris et al. (2016) argued that situational strength of SJT items – and specifically the clarity of SJT items – should moderate the relation of personality and SJT performance (see also; Campion & Ployhart, 2013; Martin-Raugh & Kell, 2019; Rockstuhl & Lievens, 2020). That is, no clear directive for appropriate behavior may be given in ambiguous situations. Test-takers may rely completely on their own trait dispositions to respond to SJT items (Harris et al., 2016). In fact, personality traits have been demonstrated as main antecedents of SJT performance (see McDaniel et al., 2007). However, this relation may vary depending on the situational strength of situations in SJT items. That is, SJT items that are high in situational strength may show a lower correlation with personality traits than SJT items that are low in situational strength? This assumption relies on a conceptualization of SJT items as simulations of work-related scenarios, similar to assessment center exercises, in which situation descriptions are essential for underlying psychological processes (e.g., McDaniel & Nguyen, 2001; Motowidlo et al., 1990; Weekley et al., 2015).

Despite these arguments, the situational influences on SJT performance are subject to debate. Recently, Lievens and Motowidlo (2016) reconceptualized SJTs as measure of general domain knowledge. They argued that responses to SJTs are not situation-specific and that test-takers use general beliefs about the utility of trait-related behaviors to respond to SJT items. These so-called implicit trait policies are defined to capture individual beliefs about the effectiveness of behaviors that reflect a specific trait (Motowidlo et al., 2006a, 2006b; Motowidlo & Beier, 2010). Implicit trait policies are developed from personality trait dispositions and general life experiences and are thus not dependent on specific job experiences or knowledge about effective behaviors in specific situations (Motowidlo et al., 2006a, 2006b; Motowidlo & Beier, 2010). This reconceptualization of SJTs specifically builds on results that showed, for a large number of SJT items, no differences in item difficulty when situation descriptions were omitted (Krumm et al., 2015). Krumm and colleagues argued that situation descriptions may be less relevant for SJT performance than conceptually defined and that some other processes may take place. In fact, a series of recent further showed no or only little relevance

of situation descriptions for the construct-related validity of SJTs (Schäpers et al., 2019; Schäpers, Freudenstein, et al., 2020) and similar effects on item difficulties for a video-based SJT (Schäpers, Lievens, et al., 2020).

Two recent studies took a closer look at the key positions of the debate about the underlying processes of SJT performance. The first study demonstrated that several measures of implicit trait policies lacked construct-related validity (Freudenstein, Mussel, et al., 2020). As much more variance was test-specific than shared among measures, these results challenged the notion of SJTs as context-independent measures. The second study suggested that the debate about whether or not situations are relevant for SJT performance has so far neglected complex relations between situation cues in SJT items and psychological relevant situations (Freudenstein, Schäpers, et al., 2020; see also Brown et al., 2016). These authors demonstrated that situation construal of SJT items predicted response behavior in SJT items regardless of whether the situation description was presented or not. They concluded that response option contain sufficient situation cues to construe psychologically relevant situations (see also Harris et al., 2016; Melchers & Kleinmann, 2016). Overall, these results speak in favor of an interactionist perspective of person and situations on SJT items similar to psychological processes in other simulation-based selection methods (e.g., assessment centers; Jansen et al., 2013).

Situational strength is one of the most regarded concepts to explain situational influences in the prediction of job performance. So, assessing the relevance of SJT items' situational strength for the relation of personality traits SJT performance is needed to further understand whether SJTs measure person-situation processes. Moreover, such knowledge would contribute to an understanding of why SJTs predict job performance criteria (see Christian et al., 2010; McDaniel et al., 2001, 2007). However, to the best of our knowledge, the influence of situational strength on the relation of personality and SJT performance has not been directly tested. In this study, we do so by assessing the situational strength of several SJT item. Against the background of the current debate on the relevance of situation cues for SJT responses (e.g., Freudenstein, Schäpers, et al., 2020; Krumm et al., 2015; Lievens & Motowidlo, 2016), we refrain from postulating a specific hypothesis:

RQ: Does the situational strength of SJT items negatively moderate the relation of broad personality traits and SJT responses?

Methods

In this study, we reanalyzed three data sets that were previously reported by Freudenstein et al. (2020; data is available on the Open Science Framework; osf.io/6kd9h) and Schäpers et al. (2019; data was provided by the first author). These data contain four different work-related SJTs with a total of 44 SJT items and self-reported Big Five personality. For some studies, SJT items were experimentally manipulated between subjects (e.g., omitting situation descriptions; Schäpers et al., 2019). Hence, we only included data from participants who completed the unmanipulated versions of the SJT items. We further considered participants eligible for this study if complete personality data were available. Importantly, we directly assessed situational strength of SJT items in addition to these

data. This was done by collecting subject matter expert ratings of situational strengths facets for all SJT items.

Sample

Overall, 718 participants from the previous studies were eligible for this study (n1 = 104; n2 = 315; see Schäpers et al., 2019; n3 = 299; see Freudenstein, Schäpers, et al., 2020). We tested the data for careless responding by computing Mahalanobis distances for the self-reported personality data (Meade & Craig, 2012). Based on an $\alpha = .001$ criterion, we excluded $n = 11$ participants from further analyses. Thus, we analyzed a total sample of $N = 707$ (n_1 =101; n_2 =313; n_3 =293; 451 female). On average, the sample was 32.87 (SD = 13.38; range: 18-78) years old. For detailed descriptions of sample characteristics and data collection see Schäpers, Mussel, et al. (2019) and Freudenstein et al. (2020).

Measures

Team Role Test

The Team Role Test (TRT; Mumford et al., 2008) is a 10-item SJT that assesses knowledge about suitable team roles in specific situations. The data set provided by Schäpers et al. (2019) included a total of 313 participants, who responded to a modified version of the TRT, which comprised four response options instead of 10 for each item. Test-takers were asked what they should do in each situation. The most effective response option of each situation was scored as "1". All other response options were scored as "-1". Reliability of this test was low, but within the meta-analytical range of SJT's reliability ($\omega = .28$; $\alpha = .41$; Catano et al., 2012; Kasten & Freund, 2016).

Situational Judgment Questionnaire

The Situational Judgment Questionnaire (SJQ; Motowidlo et al., 2006b) consists of 22 items with four response options. The test assesses work-related behavior in the presence of other people such as supervisors or coworkers. In this SJT, all response options are designed to express agreeableness (Motowidlo et al., 2006b). The data set by Schäpers et al. (2019) comprised 10 items of this SJT, which asked participants how they should behave in each situation ($n = 313$). Effective response options were scored as "1" and ineffective response options were scored as "-1". Reliability of this SJT was $\omega = .41$ and $\alpha = .59$.

Personal Initiative SJT

The Personal Initiative SJT (PI-SJT; Bledow & Frese, 2009) consists of 12 situation descriptions with four to five response options. It assesses personal initiative in work-related settings. This SJT was applied to two out of the three samples ($n = 394$). Schäpers, Mussel et al. (2019) asked participants how they would behave in each situation. However, Freudenstein, Schäpers et al. (2020) additionally asked participants how they would not act in each situation. For consistency, only responses to the question "what would you do" were considered for these analyses. Effective response options were scored as "1", ineffective response options as "-1", and all remaining response options as "0". Reliability of this SJT was $\omega = .62$ and $\alpha = .68$. As we used data from two samples for this SJT, we tested for measurement invariance between the two samples. The general factor model had good model fit, $\chi^2(54) = 69.393$, $p = .077$; CFI = .957; RMSEA = .027; SRMR = .060. Further, factor loadings could be restrained to equality between samples without a decrease in

model fit (i.e., metric invariance), $\Delta\chi^2(11) = 6.022$, $p = .872$; ΔCFI = -.021; ΔRMSEA = .006.

Teamwork SJT

The Teamwork SJT (TW-SJT Gatzka & Volmer, 2017) measures effective behavior in teamwork situations. It consists of 12 situation descriptions with four response options. Depending on the situation descriptions, participants (n = 104; Schäpers et al., 2019) were asked how they should behave or what their team should do. Effective response options were scored as "1", ineffective response options as "-1", and all remaining response options as "0". Reliability of this SJT was $\omega = .53$ and $\alpha = .62$.

Self-Reported Personality

All 718 participants responded to the German short version of the Big Five Inventory (BFI-K; Rammstedt & John, 2005). This inventory consists of 21 items assessing the broad personality traits emotional stability, extraversion, openness, agreeableness, and conscientiousness. Participants were asked to indicate on a 5-point rating scale (1 = disagree strongly; 5 = agree strongly) whether each item described themselves appropriately. We tested whether this scale showed metric measurement invariance (i.e., identical factor loadings) across the three different samples. However, one item from the Openness factor failed this test. Hence, we removed this item from all analyses. Since broad personality measures typically do not meet fit criteria for latent models (see Hopwood & Donnellan, 2010), the model fit of the resulting 20-item scale can be interpreted as acceptable; $\chi^2(160) = 706.581$, $p < .001$; CFI = .859; RMSEA = .070; SRMR = .069. Reliabilities for the five factors ranged from $\omega = .68$ to .83 and $\alpha = .67$ to .82 (see Table 1). Importantly, model fit did not differ when factor loadings were restrained to equality across the three different samples; $\Delta\chi^2(30) = 32.914$, $p = .326$; ΔCFI = .001; ΔRMSEA = .002.

Situational Strength

Two authors and one research assistant with particular expertise in SJT research independently evaluated the situational strength of all 44 SJT items. To do so, we used three items with the highest item-total correlation of each factor from the job-related situational strength questionnaire (Meyer et al., 2014). This measurement comprises four factors, namely Clarity (e.g., "specific information about work-related responsibilities is provided"), Consistency ("different sources of work information are always consistent with each other"), Constraints (e.g., "procedures prevent an employee from working in his/her own way"), and Consequences (e.g., "mistakes are more harmful than they are for almost all other situations"). Since SJT items typically do not contain enough situational context to assess the situation's consistency, we assessed situational strength only on the remaining three scales. Importantly, we instructed raters to take the situation description and response options into consideration, as previous research suggested that relevant situation cues may also be present in response options (Freudenstein, Schäpers, et al., 2020; Harris et al., 2016; Melchers & Kleinmann, 2016). Internal consistency for the three factors ranged from $\alpha = .84$ to .90[11]. We computed mean scores for each factor within raters. Initial inter-item correlation (ICC3,k) for these scores ranged from .63 to .79 thus indicating moderate to strong interrater agreement (LeBreton & Senter,

[11] Note that these values are based on a sample size of $n = 3$. However, Meyer et al. (2014) reported very similar values. Due to the small sample size, coefficients ω was not computed.

Table 1

Descriptive Statistics, Bivariate Correlations, and Internal Consistencies

	M (*SD*)	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. ES	3.13 (0.92)	*(.79)*								
2. E	3.58 (0.85)	.23*	*(.82)*							
3. O	4.07 (0.73)	.03	.26*	*(.67)*						
4. A	3.08 (0.82)	.22*	.13*	.09*	*(.67)*					
5. C	3.77 (0.69)	.19*	.17*	.09*	.08*	*(.70)*				
6. PI	1.40 (3.74)	.15*	.25*	.15*	.06	.16*	*(.68)*			
7. TRT	1.75 (3.07)	.01	.04	.03	.12*	.10	-	*(.41)*		
8. TW	5.23 (2.88)	.22*	.02	-.03	.08	.19	.40*	-	*(.61)*	
9. SJQ	3.64 (3.35)	.03	.03	.00	.09	.02	-	.31*	-	*(.59)*

Notes. n = 87 – 707. Coefficient omega in parentheses on diagonal. As not all participants responded to all SJTs, some bivariate correlations among SJT scores could not be computed. ES = Emotional Stability, E = Extraversion, O = Openness, A = Agreeableness, C = Conscientiousness, PI = Personal Initiative SJT, TRT = Team Role Test, TW = Teamwork SJT, SJQ = Situational Judgment Questionnaire. * $p < .05$.

2008). Hence, we collapsed ratings from all three raters.

Data Analyses

To test our research question, we examined an interaction effect of situational strength and personality on SJT responses. To analyze all data simultaneously, we combined all three data sets and fitted ordinal mixed-effects models with crossed random effects (Baayen et al., 2008; Tutz & Hennevogl, 1996) using the R package *ordinal* (Christensen, 2018). This combination of data was possible, as missing data only occurred on the dependent variable (i.e., SJT responses), which should not lead to biases in regression coefficients (Little, 1992). We included random intercepts for SJT items and persons to appropriately account for systematic variance components in SJT responses due to individuals and item content. We fitted separate models for each Big Five trait by stepwise including personality as person-level predictor of SJT response, random slopes for personality across SJT items, the three facets of situational strength as item-level predictor, and the interaction of the item-level and person-level predictors. Following guidelines by Enders and Tofighi (2007), we scaled situational strength variables within SJT items and personality variables within individuals. The R code is available on the Open Science Framework (osf.io/hck3j).

Results

Descriptive statistics of SJT performance and personality are depicted in Table 1. The Shapiro-Wilk test for normality showed that all situational strength facets were approximately normal distributed (W = 0.976, p = .470). Importantly, situational strength in SJT items ranged from weak to strong. That is, we found

substantial variability in all three facets (see Table 2; $range_{Clarity}$: 1.78-6.67; $range_{Constraints}$: 1.78-6.00; $range_{Consequences}$: 1.33-6.22), thus enabling us to observe possible moderating effects. Moreover, clarity as well as consequences were positively correlated with SJT item difficulty (*r*s = .15 and .24) and the relative frequency of the most chosen response option (*r*s = .19 and .24). This indicates that the situational strength had a restricting effect on individual responses in that SJT items became easier with increasing situational strength. However, due to the small sample size (n = 44) no statistical significance was reached (*p*s = .110-.322). Conversely, constraints were negatively correlated with SJT item difficulty and the relative frequency of the most chosen response option (*r*s = -.10 and -.13, *p*s = .534 and .402).

Next, we found that all Big Five traits significantly predicted SJT performance, although effect sizes were small (*B*s = 0.04-0.05, *p*s = .007-.046). When we included random slopes to the model, the fit only increased for conscientiousness. Thus, the prediction of SJT performance across items differed only for conscientiousness; $\Delta\chi^2(2) = 8.578$, $p = .014$. The inclusion of clarity, constraints, and consequences as fixed effects did not increase model fit. Finally, tests of interactions between facets of situational strength and personality produced mixed findings. For all five traits, likelihood-ratio tests indicated that models with interaction terms did not differ from models without interaction terms; $\Delta\chi^2(3) = 0.34$-6.79, *p*s = .079-.709. However, the facet constraints interacted positively with emotional stability ($B = 0.03$, $p = .035$) and extraversion ($B = 0.04$, $p = .009$). This indicates that more situational constraints led to an increased relation between emotional stability or extraversion, respectively, with SJT performance.

Discussion

This study sought to examine the relevance of SJT items' situational strength on SJT performance and, more specifically, situational strength as a moderator on the relation of personality and SJT performance. First, we found that all Big Five traits significantly predicted SJT performance. This result is in line with previous meta-analytical findings revealing personality as main antecedent of SJT performance (McDaniel et al., 2007). Second, situational strength of SJT items had no significant direct effect on SJT performance. That is, test takers did not score significantly higher on items that were high in situational strength, although

Table 2

Descriptive Statistics, Bivariate Correlations, and Internal Consistencies of Situational Strength Variables

	M (*SD*)	1.	2.	3.
1. Clarity	4.65 (1.14)	*(.63)*		
2. Constraints	4.10 (1.21)	.26	*(.65)*	
3. Consequences	3.12 (1.15)	.56*	.11	*(.79)*

Notes. n = 44. ICCs in parentheses on diagonal.

preliminary evidence suggests that variability in responses was reduced for stronger SJT items. Third, we found no moderating effect for most situational strength facets on the relationship between personality traits and SJT item responses. Notably, we did not find a moderated link between conscientiousness and SJT performance. This is surprising as effects of situational strength are well established for the link between conscientiousness and job performance (e.g., Meyer et al., 2009), as well as conscientiousness being the most relevant antecedent of SJT performance (McDaniel et al., 2007). Furthermore, we did not find moderation effects for the situational strength facet clarity. Clarity has been proposed as the most relevant situational strength facet in the context of SJTs (Harris et al., 2016). In sum, we conclude that situational strength of SJT items does not moderator the relationship between personality and SJT performance.

Overall, these results support the emerging notion of SJTs as context-independent measures (e.g., Krumm et al., 2015; Lievens & Motowidlo, 2016). Most relevant to our results, situation descriptions had no influence on the correlation of SJT responses and personality traits (Schäpers et al., 2019; Schäpers, Freudenstein, et al., 2020). The authors argued that trait-activating processes in SJT items may not solely be due to situation descriptions. Note that trait-activation is a necessary condition for effects of situational strength (Tett & Burnett, 2003). Only if situations are relevant for a specific trait, the strength of a situation may moderate the relation of that trait and behavior. Hence, the current study further suggests that situation descriptions in SJT item do not evoke situational processes that are similar to processes underlying real-life behavior. Rather, responses to SJT items may reflect general ratings of response options. That is, SJT performance may correlate with personality traits, as rating response options represents a similar task compared to self-report questionnaires (see Schäpers, Freudenstein, et al., 2020).

Surprisingly, the facet constraints positively moderated the relation of emotional stability and extraversion, respectively, with SJT performance. These effects were in the opposite direction of situational strength theory as stronger situations (i.e., more situational constraints) was associated with a stronger link between personality and SJT performance. However, more situational constraints in SJT items did not reduce variability in response behavior although this reflects an essential condition of situational strength theory (see Keeler et al., 2019). Generally, such mixed effects with regard to the influence of situational strength on the response variability among individuals are not unique to this study. For instance, Meyer et al. (2014) found that the negative link between conscientiousness and counterproductive work behavior was more pronounced in strong situations. negatively moderated the relation of conscientiousness and counterproductive work behavior. These authors attributed these findings to complex processes of how individuals perceive situational strength as psychologically relevant.

Psychologically relevant perceptions of situations have also been shown to predict SJT performance (Freudenstein, Schäpers, et al., 2020). This speaks in favor of the situation-dependent view on SJTs, which – in the current study – was only supported by two significant moderation effects. Notably, Freudenstein and colleagues (2020) also found that relevant situation perceptions can be evoked by

response options and that situation descriptions are negligible for theses processes. Overall, the format of fixed behavioral response options in SJT items may hinder the full potential of underlying situational processes of SJT responses. This may also be true for situational strength, as behavioral response options in SJT items may not reflect a broad range of trait-related behaviors and, as a result, test-takers may be unable to respond consistently with their personality for weak or moderately strong situations. Moreover, behavioral response options may not always reflect those behaviors that are demanded by stronger situations so that variability in responses may emerge even though situations per se are strong.

In view of these results more research is needed to further examine processes of person-situation interactions underlying SJT performance. Situational strength has been particularly useful to understanding when and how personality predicts job performance (e.g., Meyer et al., 2009). These processes should be taken into consideration when developing SJTs and other tools for personnel selection. However, different approaches are needed to enhance our understanding of situational processes in simulation-based assessments. Experimental test validation (Krumm et al., 2019) may be particularly useful to examine specific processes. For instance, researches may design specific situation descriptions that vary in situational strength and align response options to reflect trait-related behavior that corresponds with the situation descriptions. In addition, recent research on situation taxonomies may be useful to understand how short situation descriptions may evoke meaningful perceptions (e.g., Rauthmann et al., 2014; Saucier et al., 2007; see also Lievens, 2017a).

Given these uncertainties about the relevance of person-situation interactions for SJT performance, we urge practitioners to rely on construct-driven SJTs (Guenole et al., 2017; Lievens, 2017b). Although the underlying principles with regard to the role of situation descriptions do not differ from traditional SJTs (Schäpers, Freudenstein, et al., 2020), these tests validly assess unidimensional constructs (e.g., Mussel et al., 2018; Olaru et al., 2019; Oostrom et al., 2018). Situation descriptions in construct-driven SJTs may function as a highly specific frame-of-reference, which may enhance the criterion related validity (see Shaffer & Postlethwaite, 2012).

Limitations

We assessed situational strength on the item level. This limits the generalizability of our results in two ways. First, we analyzed only 44 SJT items. This number is slightly lower than the typical recommendation of at least 50 level two units for unbiased estimates of standard errors (see Maas & Hox, 2005). That is, possible effects of situational strength may have been undetected in our study, especially if true effects are small. Second, our results are based upon the assumption that an objective situational strength of SJT items exist. Nevertheless, situational strength may also be understood as some sort of situation construal (see Meyer et al., 2014). That is, the individual perception of situational strength may matter in order to influence behavior. Future research should incorporate this perspective on situational strength and ask individuals about their perception of situational strength in SJT items.

We also specifically tested work-related SJTs. These SJTs are designed to reflect multidimensional behavioral

tendencies rather than specific constructs (McDaniel et al., 2007, 2016; Weekley et al., 2015). Accordingly, we assumed that all SJT items had at least some relevance for a broad range of personality traits. Although we found significant effects for all personality traits, nuances in the trait-activating potential of SJT items may exist (see Tett & Burnett, 2003). For instance, we found significant random slopes for the prediction of SJT performance by conscientiousness. This may reflect that some SJT items were more trait-activating for conscientiousness than others. Moreover, fixed effects of personality traits on SJT performance were rather small, which may further add to the notion that not all items were equally relevant to all personality traits. So, we propose that further examinations of situational strength in the context of SJTs specifically take the trait-activation potential into account. Construct-driven SJTs may pose as a good starting point for such an undertaking (see Guenole et al., 2017).

Conclusion

In this study, we built on situational strength theory to examine whether SJT performance reflected person-situation processes. Similar to the influence of situational strength on the relation of conscientiousness and job performance, we argued that situations in SJTs items may have the same underlying mechanism (Harris et al., 2016). However, our results demonstrated that this may not be the case. This study shows that the debate about underlying psychological processes of SJT performance is not yet resolved. Whereas some person-situation processes may be relevant to SJT responses (i.e., situation construal; Freudenstein, Schäpers, et al., 2020), others (i.e., situational strength) may be not. Overall, we call for more theory-driven SJT developments that provide clear and verifiable assumptions about underlying psychological processes (see Guenole et al., 2017).

References

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*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bledow, R., & Frese, M. (2009). Situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, *62*(2), 229–258. https://doi.org/10.1111/j.1744-6570.2009.01137.x

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*(1), 38–42. https://doi.org/10.1017/iop.2015.113

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures. In Neil D. Christiansen & Robert P. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). Routledge.

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*(3), 333–346. https://doi.org/10.1111/j.1468-2389.2012.00604.x

Christensen, R. H. B. (2018). *Cumulative link models for ordinal regression with the R package ordinal* [Manuscript submitted for publication]. https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf

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*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Christiansen, N. D., Hoffman, B. J., Lievens, F., & Speer, A. B. (2013). Assessment centers and the measurement of personality. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 477–497). Routledge. https://doi.org/10.13140/2.1.3105.9847

Dalal, R. S., Meyer, R. D., Bradshaw, R. P., Green, J. P., Kelly, E. D., & Zhu, M. (2014).

Personality strength and situational influences on behavior. *Journal of Management*, *41*(1), 261–287. https://doi.org/10.1177/0149206314557524

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

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*(4), 1667–1684. https://doi.org/10.1111/j.1751-9004.2008.00122.x

Freudenstein, J.-P., Mussel, P., & Krumm, S. (2020). *On the construct-related validity of implicit trait policies* [Manuscript in prepared for publication].

Freudenstein, J.-P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. https://doi.org/10.1111/peps.12385

Gatzka, T., & Volmer, J. (2017). Situational Judgment Test für Teamarbeit (SJT-TA) [situational judgment test for teamwork]. In *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis249

Guenole, N., Chernyshenko, O. S., & Weekly, J. A. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, *17*(3), 234–252. https://doi.org/10.1080/15305058.2017.1297817

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*(1), 23–28. https://doi.org/10.1017/iop.2015.110

Herde, C. N., & Lievens, F. (2018). Multiple speed assessments. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000512

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Pers Soc Psychol Rev*, *14*(3), 332–346. https://doi.org/10.1177/1088868310361240

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*(2), 326–341. https://doi.org/10.1037/a0031257

Judge, T. A., & Zapata, C. P. (2015). The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance. *Academy of Management Journal*, *58*(4), 1149–1179.

Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). *European Journal of Psychological Assessment*, *32*(3), 230–240. https://doi.org/10.1027/1015-5759/a000250

Keeler, K. R., Kong, W., Dalal, R. S., & Cortina, J. M. (2019). Situational strength interactions: Are variance patterns consistent with the theory? *Journal of Applied Psychology*, *104*(12), 1487–1513. https://doi.org/10.1037/apl0000416

Krumm, S., Hüffmeier, J., & Lievens, F. (2019). Experimental test validation: Examining the path from test elements to test performance. *European Journal of Psychological Assessment*, *35*(2), 225–232. https://doi.org/10.1027/1015-5759/a000393

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology*, *100*(2), 399–417. https://doi.org/10.1037/a0037674

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lee, S., & Dalal, R. S. (2016). Climate as situational strength: Safety climate strength as a cross-level moderator of the relationship between conscientiousness and safety behaviour. *European Journal of Work and Organizational Psychology*, *25*(1), 120–132.

Lievens, F. (2017a). Assessing personality-situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*(5), 424–440. https://doi.org/10.1002/per.2111

Lievens, F. (2017b). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing*, *17*(3), 269–276. https://doi.org/10.1080/15305058.2017.1309857

Lievens, F., & De Soete, B. (2012). Simulations.

In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 383–410). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199732579.013.0017

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*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology*, *100*(4), 1169–1188. https://doi.org/10.1037/apl0000004

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao (Eds.), *Research in personnel and human resources management* (pp. 99–152). Emerald Group Publishing Limited. https://doi.org/10.1108/S0742-7301(2009)0000028006

Little, R. J. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*(420), 1227–1237.

Maas, C., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-1881.1.3.86

Martin-Raugh, M. P., & Kell, H. J. (2019). A process model of situational judgment test responding. *Human Resource Management Review*. https://doi.org/10.1016/j.hrmr.2019.100731

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*(1), 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: Perspectives on Science and Practice*, *9*(1), 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*(4), 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*(1-2), 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*(3), 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*(1), 29–34. https://doi.org/10.1017/iop.2015.111

Melchers, K. G., Wirz, A., & Kleinmann, M. (2012). Dimensions AND exercises: Theoretical background of mixed-model assessment centers. *The Psychology of Assessment Centers*, 237–254.

Meyer, R. D., Dalal, R. S., & Bonaccio, S. (2009). A meta-analytic investigation into the moderating effects of situational strength on the conscientiousness–performance relationship. *Journal of Organizational Behavior*, *30*(8), 1077–1102.

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*(1), 121–140. https://doi.org/10.1177/0149206309349309

Meyer, R. D., Dalal, R. S., José, I. J., Hermida, R., Chen, T. R., Vega, R. P., Brooks, C. K., & Khare, V. P. (2014). Measuring job-related situational strength and assessing its interactive effects with personality on voluntary work behavior. *Journal of Management*, *40*(4), 1010–1041. https://doi.org/10.1177/0149206311425613

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, *80*, 252–283.

Mischel, W. (1977). The interaction of person and situation. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333–352). Lawrence Erlbaum.

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*(2), 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*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L.

(2006a). 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). Lawrence Erlbaum Associates.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*(4), 749–761. https://doi.org/10.1037/0021-9010.91.4.749

Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The team role test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology*, *93*(2), 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*(5), 328–335. https://doi.org/10.1027/1015-5759/a000346

Olaru, G., Burrus, J., Maccann, C., Zaromb, M. F., Wilhelm, O., & Roberts, D. R. (2019). Situational judgment tests as a method for measuring personality: Development and validity evidence for a test of dependability. *PLoS One*, *14*(2), e0211884. https://doi.org/10.1371/journal.pone.0211884

Oliver, T., Hausdorf, P., Lievens, F., & Conlon, P. (2016). Interpersonal dynamics in assessment center exercises: Effects of role player portrayed disposition. *Journal of Management*, *42*(7), 1992–2017. https://doi.org/10.1177/0149206314525207

Oostrom, J. K., de Vries, R. E., & de Wit, M. (2018). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*(1), 1–29. https://doi.org/10.1080/08959285.2018.1539856

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K). *Diagnostica*, *51*(4), 195–206. https://doi.org/10.1026/0012-1924.51.4.195

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677–718. https://doi.org/10.1037/a0037250

Rockstuhl, T., & Lievens, F. (2020). Prompt-specificity in scenario-based assessments: Associations with personality versus knowledge and effects on predictive validity. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000498

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*(3), 479–503. https://doi.org/10.1111/j.1467-6494.2007.00446.x

Schäpers, P., Freudenstein, J.-P., Mussel, P., Lievens, F., & Krumm, S. (2020). Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests. *Journal of Research in Personality*. https://doi.org/10.1016/j.jrp.2020.103963

Schäpers, P., Lievens, F., Freudenstein, J.-P., Hüffmeier, J., König, C. J., & Krumm, S. (2020). 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*, *93*(2), 472–494. https://doi.org/10.1111/joop.12297

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

Schollaert, E., & Lievens, F. (2012). Building situational stimuli in assessment center exercises: Do specific exercise instructions and role-player prompts increase the observability of behavior? *Human Performance*, *25*(3), 255–271. https://doi.org/10.1080/08959285.2012.683907

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*(3), 445–494. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Tett, & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*(3), 500–517. https://doi.org/10.1037/0021-9010.88.3.500

Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics & Data Analysis*, *22*(5), 537–557. https://doi.org/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*(1), 295–322. https://doi.org/10.1146/annurev-orgpsych-032414-11130

Chapter 5

Standardized State Assessment: A Methodological Framework to Assess Person-Situation Processes in Hypothetical Situations

This article has been prepared for publication:

Freudenstein, J.-P., Schulze, J., Schäpers, P., Mussel, P., & Krumm, S. (2020). *Standardized state assessment: A methodological framework to assess person-situation processes in hypothetical situations* [Manuscript prepared for publication].

Standardized State Assessment: A Methodological Framework to Assess Person-Situation Processes in Hypothetical Situations

Jan-Philipp Freudenstein
Freie Universität Berlin

Julian Schulze
Freie Universität Berlin

Philipp Schäpers
Singapore Management University

Patrick Mussel, & Stefan Krumm
Freie Universität Berlin

In response to contemporary personality theories, psychological assessments are increasingly concerned with person-situation processes. Most commonly, ambulatory assessments sample individuals within their real-life environments, but further approaches exist that aim at measuring person-situation processes by incorporating hypothetical situation descriptions. Thus far, no common guidelines exist on how to develop such measures so that they validly assess person-situation processes. In this article, we propose Standardized State Assessment as a methodological framework for the assessment of situation-specific states in hypothetical situations. We build on theoretical advances in personality research and previous assessment approaches to derive guidelines for a theory-driven development of hypothetical situation descriptions. We further describe how states should be measured in these situations. Finally, we propose that appropriate latent measurement models and validation strategies may help to develop assessments that are similar to real-life person-situation processes. Standardized State Assessment may offer economically advantageous alternatives for research that is unable to adopt ambulatory assessments. Moreover, we discuss whom this framework may help to answer theoretical questions on person-situation processes.

Keywords: Person-situation processes, State Measures, Personality Assessment

Overcoming the person-situation debate, researchers came to an agreement that both person and situation processes influence individual behavior (e.g., Fleeson & Noftle, 2008; Mischel & Shoda, 1998). Hence, Whole Trait Theory integrates individual dispositions and situational influences by positing that personality is best described as a density distribution of momentary states of behavior, thoughts, and feelings (i.e., descriptive part of traits; Fleeson, 2001;

Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019). Consistent tendencies in behavior, thoughts, and feelings are reflected in means of individual state distributions (e.g., Epstein, 1979; Fleeson, 2001; Fleeson & Gallagher, 2009; Jones et al., 2017; Rauthmann et al., 2019). Further, variability of states around that mean comprises valuable information about individuals' personality as well (e.g., Bleidorn, 2009; Fleeson, 2001, 2007; Fleeson & Jayawickreme, 2015; Heller et al., 2007; Jones et al., 2017; Rauthmann et al., 2016; Wilson et al., 2017). Importantly, Whole Trait Theory emphasizes that personality theories must strive to explain complete state distributions beyond an individual's central tendency in behavior, thoughts and feelings, and thus need to move beyond the focus of existing trait theories (i.e., explanatory part of traits; Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019).

The most commonly used methodology to assess state distributions and person-situation processes in general is ambulatory assessment (e.g., Bleidorn, 2009; Fleeson, 2007; Heller et al., 2007; Rauthmann et al., 2016; Wilson et al., 2017). Ambulatory assessment approaches measure psychological constructs or behavior repeatedly over time (Fisher & To, 2012; Hofmans et al., 2019; Shiffman et al., 2008; Trull & Ebner-Priemer, 2013, 2020) and are thus "uniquely suited to focus on within-individual processes" (Trull & Ebner-Priemer, 2020, p. 56). A key advantage of such momentary assessments is the reduction of retrospective biases as individuals report feelings, thoughts, or behaviors as they occur in a given situation (Trull & Ebner-Priemer, 2020). Furthermore, measurement precision increases due to the repeated assessment of individuals and thus a more fine-grained picture of dynamic intraindividual processes emerges (Wright & Zimmermann, 2019). Finally, ambulatory assessment enhances the ecological validity of measurements due to sampling of individuals in their natural environment and in real-life situations (Shiffman et al., 2008).

Despite these advantages, however, there is an inherent drawback when sampling individuals in their natural environment: the lack of control over sampled situations. Hence, comparisons of specific situations either across or within individuals are not possible (Fleeson & Law, 2015; Rauthmann, Sherman, Nave, et al., 2015). To disentangle effects of persons, situations, and person-situation interactions in state assessments, research designs are required, in which all participants experience identical situations (Rauthmann, Sherman, Nave, et al., 2015). To do so, we propose a methodological framework of Standardized State Assessment (SSA) that enables researchers to control situations when assessing dynamic intraindividual processes.

A key assumption of our framework is that similar psychological processes as in real-life can be evoked by hypothetical situations. Although situation vignettes and similar approaches have been used in prior research to assess dynamic intraindividual processes (e.g., Aguinis & Bradley, 2014; Blum et al., 2018; Kammrath et al., 2005; Lievens, 2017a; Lievens et al., 2018; Rauthmann, 2012; Ziegler et al., 2019), no common framework exists that guides researchers in the development of vignettes and in the evaluation of responses to these hypothetical situations to specifically measure psychological states. Such a lack of guidelines may lead to a reduced comparability between

measurements and may pose as a threat to validity. In the following, we describe how hypothetical situation descriptions should be developed to enable a more standardized and controlled capturing of meaningful intraindividual differences across situations. In doing so, we build on previous research on assessment techniques using hypothetical situation descriptions and incorporate recent insights on psychologically relevant situations (e.g., Rauthmann et al., 2014; Saucier et al., 2007). Finally, we draw on Latent State Trait Theory (LST; Steyer et al., 1992, 1999) and propose specific latent variable models (e.g., the recently introduced bifactor S-1 model; Eid et al., 2017) to disentangle trait-specific and state-specific variance components in SSAs.

Standardized State Assessment

SSA aims at providing a methodological framework for the assessment of person-situation processes that controls for situational content within and between subjects. We posit that SSA comprises two core features. First, SSA contains standardized, theory-driven situation descriptions that are designed to mimic real-life situations as closely as possible to facilitate the assessment of personality states. This feature distinguishes SSA from ambulatory assessment as SSA samples states in hypothetical situations instead of real-life situations. Second, participant responses are also gauged in a standardized way across situation descriptions. That is, participants respond to the same state measure for each situation description. All items of the state measure thereby represent a well-defined construct. This second feature resembles the approach of state measurement in ambulatory assessment (Fleeson, 2001, 2007; Heller et al., 2007; Rauthmann et al., 2016). An example situation for a SSA consisting of these two features for the assessment of a sociability state is presented in Figure 1. Importantly, fixing situations across individuals enables (i) capturing states and traits, which can be systematically compared either (ii) within or (iii) across individuals and (iv) be used for research determining the factors that drive person-situation processes. Also, (v) latent variable models can be applied to disentangle measurement error from meaningful trait and state-specific variance components. Thus, SSA is well in line with Whole Trait Theory's call for explaining general tendencies as well as individual deviations from these tendencies in individual behavior (Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019).

Within ambulatory assessment studies, similar attempts have been made to compensate for the lack of control over sampled situations. For instance, different analytical models were proposed that account for variance in states due to different contextual situation classes (e.g., at work, with friends; Geiser et al., 2015; Nestler et al., 2018). However, we argue that SSA comes with several advantages in comparison to ambulatory assessments. First, states for the same objective situation can be compared across individuals. Similarly, repeated assessments of the same situation allow to examine intraindividual variability in these identical situations. Second, SSA enables theory-driven manipulation of situational cues. Such experimental tests are particularly useful to study person-situation contingencies (see Lievens, 2017a). For instance, Rauthmann and colleagues (2014) demonstrated that the perception of duties to fulfill was prevalent in working

situations. Researchers may use SSA to directly compare individual's situation perceptions or states in similar situations that only differ in one aspect (e.g., face-to-face meeting vs. video call). Third, SSA provides a possibility to assess states in relevant situations that only occur irregularly in real-life (e.g., when observing aggressive behavior). Finally, SSA may lower the burden to participate in comparison to ambulatory assessments. That is, no longitudinal research design is needed and a reduced selection of situations may shorten the time to complete the study.

A key assumption of SSA is that situation descriptions simulate real-life situations. The notion that short situation description evoke similar processes when compared to real-life situations is not new (e.g., Aguinis & Bradley, 2014; Kammrath et al., 2005; Lievens, 2017a; Rauthmann, 2012; Van Heck et al., 1994; Ziegler et al., 2019). One of the most prominent tools to simulate real-life situations are Situational Judgment Tests (SJTs; Motowidlo et al., 1990; Weekley et al., 2015). In addition to short situation descriptions, these tests comprise several behavioral response options, of which

Figure 1

Example Standardized State Assessment

Your new neighbor[a] invites[b] you to their birthday party. When you arrive[c] at the party[d] you realize that you don't know any other guests[e].

Put yourself in this situation. What would you do?

I would act…

	1 – Strongly disagree	2	3	4	*5 – Strongly agree*
…outgoing	☐	☐	☐	☐	☐
…sociable	☐	☐	☐	☐	☐
…quiet	☐	☐	☐	☐	☐
…shy	☐	☐	☐	☐	☐
…talkative	☐	☐	☐	☐	☐

Notes. The situation description was adapted from an SJT by Mussel et al. (2018). The state measure was adapted from the sociability scale of the BFI-2 (Soto & John, 2017).
[a] association cue, [b] process cue, [c] action cue, [d] location cue, [e] trait-activating cue.

individuals typically choose one that resembles how they would behave in a real situation (McDaniel & Nguyen, 2001). Similarly, experimental vignette methodology proposes short situation descriptions to enhance the ecological validity of experimental studies (Aguinis & Bradley, 2014). Importantly, this line of research showed that behavioral responses to short situation descriptions correlated with real-life behavior and personality traits (Blum et al., 2018; Christian et al., 2010; Lievens et al., 2018, p. 200; McDaniel et al., 2001, 2007; Ziegler et al., 2019). Moreover, responses to several situation descriptions show substantial intraindividual variability (Blum et al., 2018; Lievens et al., 2018; Rauthmann, 2012), which is related to intraindividual variability in real-life state assessments (Lievens et al., 2018). However, validity problems emerge when these methodologies are used to assess intraindividual variability. For instance, a large body of research demonstrates that situation descriptions in SJTs are less relevant for underlying processes of response measure than theory suggests (e.g., Freudenstein et al., 2020; Jackson et al., 2016; Krumm et al., 2015; Schäpers et al., 2019; Schäpers, Freudenstein, et al., 2020; Schäpers, Lievens, et al., 2020). Furthermore, correlations among intraindividual variability in an SJT and real-life states were rather low ($r \approx .20$; Lievens et al., 2018). This is partly due to the reason that no overarching framework exists that is specifically targeted the assessment of personality states with situation descriptions, although other guidelines for methods that use situation descriptions for other purposes exist (see Aguinis & Bradley, 2014; Corstjens et al., 2017).

Within the SSA framework we try to overcome these shortcomings by building on the notion that states are best assessed in real-life situations. Thus, SSA needs to mimic key characteristics of real-life situations as closely as possible. In the following, we will describe how research on situations and situation taxonomies may be incorporated into the development of situation descriptions to best mimic real-life situations. Furthermore, we will describe how states may be best assessed with hypothetical situation descriptions. Finally, we will elaborate on how appropriate measurement models, accounting for general tendencies, intraindividual differences in behavior, and measurement error, may be applied. All elements of the SSA framework are summarized in Table 1.

Situation Descriptions for Standardized State Assessment

Situational information can be described as cues, characteristics, or classes (Rauthmann, 2015; Rauthmann et al., 2014; Rauthmann, Sherman, & Funder, 2015). Cues reflect objectively observable descriptions of situation elements such as objects or persons (Rauthmann et al., 2014; Saucier et al., 2007). Characteristics represent individuals' perceptions of situations that derive from interactional processes of situation cues and personality traits (Funder, 2016; Rauthmann et al., 2014). Situation characteristics build on the notion that individual's perceptions of situations are more meaningful for predicting behavior than the objective situation cues (e.g., Mischel & Shoda, 1995; Rauthmann et al., 2014; Reis, 2008). Finally, classes categorize several situations based on similar cues or characteristics (Rauthmann et al., 2014; Rauthmann, Sherman, & Funder, 2015). In line with this classification of situations, research demonstrated that situation

characteristics predict real-life behavior (Parrigon et al., 2016; Rauthmann et al., 2014; Sherman et al., 2015), personality states (Rauthmann et al., 2016), and behavioral responses to hypothetical situation descriptions (Freudenstein et al., 2020). Accordingly, situation descriptions in SSA should contain relevant situation cues so that they evoke similar situation characteristics in individuals as do real-life situations comprising the same set of cues (see Furr & Funder, 2004).

To achieve this goal, situation descriptions should represent specific situations rather than general situation classes. That is, situation descriptions should include situation cues with a similar specificity as in situations during ambulatory assessments. In previous research, some situation descriptions tended to be quite short to the end that they only reflected rather broad situation classes (e.g., "A team member is bothered about something"; Stevens & Campion, 1994; see also Rauthmann, 2012; Ten Berge & De Raad, 2001). This type of situation descriptions leaves room for interpretation about the specifics of the situation. For example, a team member may be bothered about working conditions, a personal conflict with another colleague, or various other things. Hence, responses to these situation descriptions may either not capture specific states as individuals may aggregate their behaviors, thoughts or feelings for various applicable situations, or the assessment is not comparable as different persons may think about different situations or the same individual envisions different situations over repeated measurements (see Schulze et al., 2020). Four broad domains of cues in personality-relevant situations have been identified that may be helpful to develop precise situation descriptions (Saucier et al., 2007). These domains are locations (e.g., at work), associations (e.g., with friends, alone), actions, processes, or positions (e.g., studying or in charge), and subjective states (Saucier et al., 2007). Since subjective states will typically be the target construct of SSAs, we propose that each situation description in SSAs includes locations, associations, actions, and positions (see Figure 1).

In addition, situation descriptions should be relevant for the construct of interest (Guenole et al., 2017; Lievens, 2017b; i.e., trait-activating; see Ten Berge & De Raad, 1999, 2001, 2002; Tett & Guterman, 2000). That is, situation descriptions should give individuals the opportunity to express a behavior that is indicative for a certain personality trait rather than another trait (Tett & Guterman, 2000). The concept of trait activation is closely related to situational strength, which "denotes the *compellingness* to behave such that individual differences in behavioral dispositions are washed out" (Meyer et al., 2010; Mischel, 1977; Tett & Guterman, 2000, p. 399). Hence, the SSA should allow for variability in states: For example, a situation that is relevant for the trait of gregariousness should allow participants to express states that reflect both low and high levels of the trait (e.g., plenty of contact to other people vs. no contact to other people). This principle has already been adopted by various situation vignette techniques in personality research (e.g., Mussel et al., 2018; Oostrom et al., 2018; Rauthmann, 2012; Ziegler et al., 2019). For instance, aggregated scores of situations with trait-activating character in SJTs correlated high with self-reports of corresponding broad personality traits (Mussel et al., 2018; Olaru et al., 2019; Oostrom et al., 2018).

Table 1

Overview of the Standardized State Assessment Framework

SSA Feature	Guidelines
Situation Descriptions	- Develop specific situations instead of broad situation classes - Build on situation taxonomies to include relevant situation cues (e.g., Parrigon et al., 2016; Rauthmann et al., 2014; Saucier et al., 2007). - Include at least information about locations, associations, and actions or processes. - Allow for variability in the target construct by including trait-activating situation cues and designing situation descriptions with low to medium situational strength. - Be aware that certain situation cues may not reflect an individual's reality and that the same situation cues may be interpreted very differently by individuals. - Use information from critical incidents or ambulatory assessments to develop situation descriptions.
State Measures	- Standardize state measures across situations. - Pay attention to the validity of selected state measures. - Follow guidelines for the development of state measures (see Horstmann & Ziegler, 2020). - Use multiple indicators. - Use heuristic techniques, such as Ant Colony Optimization, to achieve valid measurement models. - Check for measurement invariance of state measures across situations.
Measurement Models	- Use latent variable models to control for measurement error. - Follow the S-1 approach to model situation-specific variance components in contrast to a reference factor (e.g., a baseline situation or a trait measure; see Eid et al., 2017). - Indicator-specific method factors may be needed to achieve good model fit.
Validity Tests	- Inspect the situation-specific latent variance. - Inspect correlations among situation-specific factors. - Inspect the convergence of state measures with a trait measure. - Compare the SSA to similar situations in a laboratory setting. - Compare the SSA to ambulatory assessment. - Examine the nomological net of the SSA. - Examine the relevance of specific situation cues with experimental test validation (see Krumm et al., 2019).

Overall, situation descriptions in SSAs should be highly specific and detailed so that participants can immerse in the exact situation. The level of immersion may be further enhanced by additional stimuli (e.g., pictures or videos; Aguinis & Bradley, 2014; Lievens & Sackett, 2006). However, with increasing level of detail in situation descriptions, two drawbacks need to be considered. First, some detailed situation cues may not be relevant or true for some individuals. To illustrate this, let us assume a situation, which describes a hypothetical interaction with the participants' sister. Such a situation presumes that all participants have a sister, which will not be true in a large number of cases. Researchers should act with caution when adopting such detailed situation descriptions. As a solution, situations with strong boundary conditions may be excluded completely. However, if the situational context is relevant to the research question, more general cues may be adopted (i.e., a family member), or only participants for which the condition is true may be eligible to participate or to answer to the particular situation.

Second, different participants will perceive identical situation cues in situation descriptions very differently in most cases. For instance, different individuals may have very different relationships to their supervisors. Hence, the cue "your supervisor" may evoke negative emotions in some participants and positive emotions in others. Researchers should generally consider whether variance in specific situation characteristics is wanted or if such cues or situations should be excluded. The idea of control for unwanted sources of variance is implemented in the experimental vignette methodology (Aguinis & Bradley, 2014). For SSA, domains of situation cues (see Saucier et al., 2007) may be manipulated experimentally across situations. Researchers may further build on taxonomies of situation characteristics to design situation descriptions (e.g., Parrigon et al., 2016; Rauthmann et al., 2014; see also Freudenstein et al., 2020; Lievens et al., 2020; Lievens, 2017a; Mussel et al., 2017). That is, situations may tap into specific domains of situation characteristics and try to control or eliminate the perception of other domains. Moreover, situation descriptions may be designed to create a conflict between opposite positions of situation characteristics. Rauthmann and colleagues (2014) mapped different dimensions of situation characteristics to relevant situation cues and relevant personality traits. For instance, the perception of Duty was positively related to situations at work and negatively related to situations with friends (see also research on dilemmas in situational interviews; Latham & Sue-Chan, 1999). Generally, experimental manipulations of situation cues or characteristics matter most when their influence on states is the focus of the research question and may be less relevant when unconditional state distributions are assessed.

Although all situation descriptions in SSAs should be developed based on theory and with regard to the research question of interest, we suggest that some kind of empirical method is used to support the development of meaningful situation descriptions. The critical incident technique is typically applied to develop scenarios in SJTs (Campion et al., 2014; Corstjens et al., 2017). Here, participants generate situation descriptions from memory, in which an individual showed very high or low expressions of the construct of interest. This technique could be enhanced by requiring participants to report cues for all relevant domains (Saucier et al., 2007)

and perceived situation characteristics (Parrigon et al., 2016; Rauthmann et al., 2014) in the reported situation descriptions. Similarly, participants in ambulatory assessment studies may be required to report this information for the occurring situation.

State Measures for Standardized State Assessment

Most commonly, states in ambulatory assessments are assessed with self-report scales that were adopted from common trait questionnaires (Fisher & To, 2012; Hofmans et al., 2019). These scales typically consist of adjective markers or rephrased items so that they are applicable to a broad range of situations. The same set of items is used to assess states for all situations, which allows for direct comparisons among different situations. Accordingly, we propose that this approach is transferred to SSAs as it maximizes the comparability between ambulatory assessments and SSAs. Although this procedure seems straightforward, it is important to explicitly incorporate it into this framework.

That is, some existing assessment methods, in which short situation descriptions are applied, use specific response options for each situation (e.g., SJTs; Weekley & Ployhart, 2006; or revealed trait technique; Costello et al., 2018). In these methods, measures to assess situation specific responses differ from situation to situation, which may induce method-specific measurement error (see Westring et al., 2009). Thus, intraindividual differences in responses to different situations may either occur due to situation-specific effects, person-situation interactions, or method-specificity in response options. In addition, recent research showed that for SJTs, participants tended to rely on situational information in response options instead of situation descriptions for their responses (Freudenstein et al., 2020). Hence, state measures should not contain situation cues and should be standardized across all situation descriptions in SSAs.

However, the validity of state measures is seldomly assessed in ambulatory assessments (Trull & Ebner-Priemer, 2020; Wright & Zimmermann, 2019). In SSAs, it is even more important that the used scales are valid for the assessment of actual states. Although situation descriptions in SSA should be developed with utmost care, they inevitably will introduce unwanted variance components as participants have to imagine how they would act, feel, or think in the given situation instead of experiencing a real situation. Thus, SSA represents approximations to real-life situations. For this reason, it is important to minimize additional sources of measurement error, such as additional contextual information in state measures. Adjective markers pose one possibility to do so (see Wiedenroth & Leising, 2020). We propose that state measures in SSA are selected based on psychometric properties in previous longitudinal studies or even specific validation studies (e.g., Zimmermann et al., 2019). Moreover, state measures should be tested for measurement invariance across situations to justify interpretations of intraindividual changes. Recent advances in the use of meta-heuristics to develop valid psychometric scales may further help to select appropriate items to assess states in SSAs (Olaru et al., 2015; Schultze, 2017). Importantly, only scales with multiple indicators for each state should be used to be able to implement the above procedures and fit latent measurement models. Recently, Horstmann and Ziegler (2020; see also

Ziegler, 2014) proposed guidelines for the development of state measures. In particular, they outlined three key questions researchers should address: what construct should be assessed, what is the purpose of the measure (e.g., what is the research question), and what is the target population (i.e., who will participate and which situations will be applied). These guidelines may serve as starting point to identify or develop suitable state measures for SSA. However, in contrast to these guidelines, state measures in SSA should exclude situation cues and complex behavioral descriptions to not confound situational information in descriptions and response options.

Measurement Models for Standardized State Assessment

Latent State Trait Theory (LST) provides a formal modelling approach to disentangling psychological state and trait variance in occasion-specific assessments (Steyer et al., 1992, 1999; Steyer & Schmitt, 1990). This approach builds on the assumption that psychological assessments are influenced by person characteristics, situation characteristics and the interaction of both (Steyer et al., 1999), which is in line with personality theories like Whole Trait Theory (Hintz et al., 2018). Hence, variance of state measures comprises reliable variance that can be divided into consistent and specific components (Steyer et al., 1992; Steyer & Schmitt, 1990). Consistency describes variance that can be attributed to interindividual differences across measurement occasions. Specificity comprises unique variance components within each measurement occasion that represent situation effects and person-situation interactions. Since LST constitutes a "generalization of classical test theory" (Steyer et al., 1999, p. 389), it allows for flexible adaptations within the structural equation modelling framework. In the simplest case, a bifactor model is specified that accounts for a general trait factor and occasion specific state residual factors. However, more complex models also exist that incorporate change processes over time, mixture distribution models (i.e., latent classes), or disentangle situation effects from person-situation interactions (see Geiser et al., 2017 for an overview). LST models are advantageous when compared to alternative estimations of trait and state variance based on manifest scale scores (e.g., computation of mean and standard deviation) as they account for measurement error of state assessments on the latent level (Geiser et al., 2017).

We propose that measurement models for SSA should generally correspond to latent state-trait models. That is, the observed variables are decomposable into measurement error, latent trait factors, and latent state residual factors. However, a key assumption of LST is that situations within individuals are randomly sampled and therefore interchangeable (Geiser et al., 2015; Steyer & Schmitt, 1990) as is typically the case in ambulatory assessment studies. This assumption justifies the simultaneous decomposition of trait variance and state residual variance in a bifactor model. However, SSA violates the interchangeability assumption of situations, as situations are predefined and no randomly sampled by the researcher. Thus, the situations as conceptualized in SSA should be considered as structurally different (i.e., situations are fixed). In such cases, typical LST models that specify bifactor structures for trait and state residual factors are not suitable (Geiser et al., 2015). As an alternative, a reference can be set to which situation specific

factors are contrasted (Eid et al., 2017; Koch et al., 2018). The model specification of state specific factors with reference follows the bifactor S-1 procedure (Eid et al., 2017). In these models a general factor across all items and additional specific factors for all situations except one are specified (see Figure 2, Panel A). Due to omitting one situation specific factor, the general factor adopts the meaning of this situation, which consists of state and trait-specific variance. Thus, the omitted situation specific factor is the reference to which all other situations can be compared. Generally, this model allows for the calculation of consistent and specific variance components between particular situations and a baseline situation (Eid et al., 2017). This model is especially useful for research designs, in which a baseline situation exists (e.g., a neutral or typical situation; Geiser et al., 2015). However, this model does not allow for differentiation of general trait and situation specific variance components.

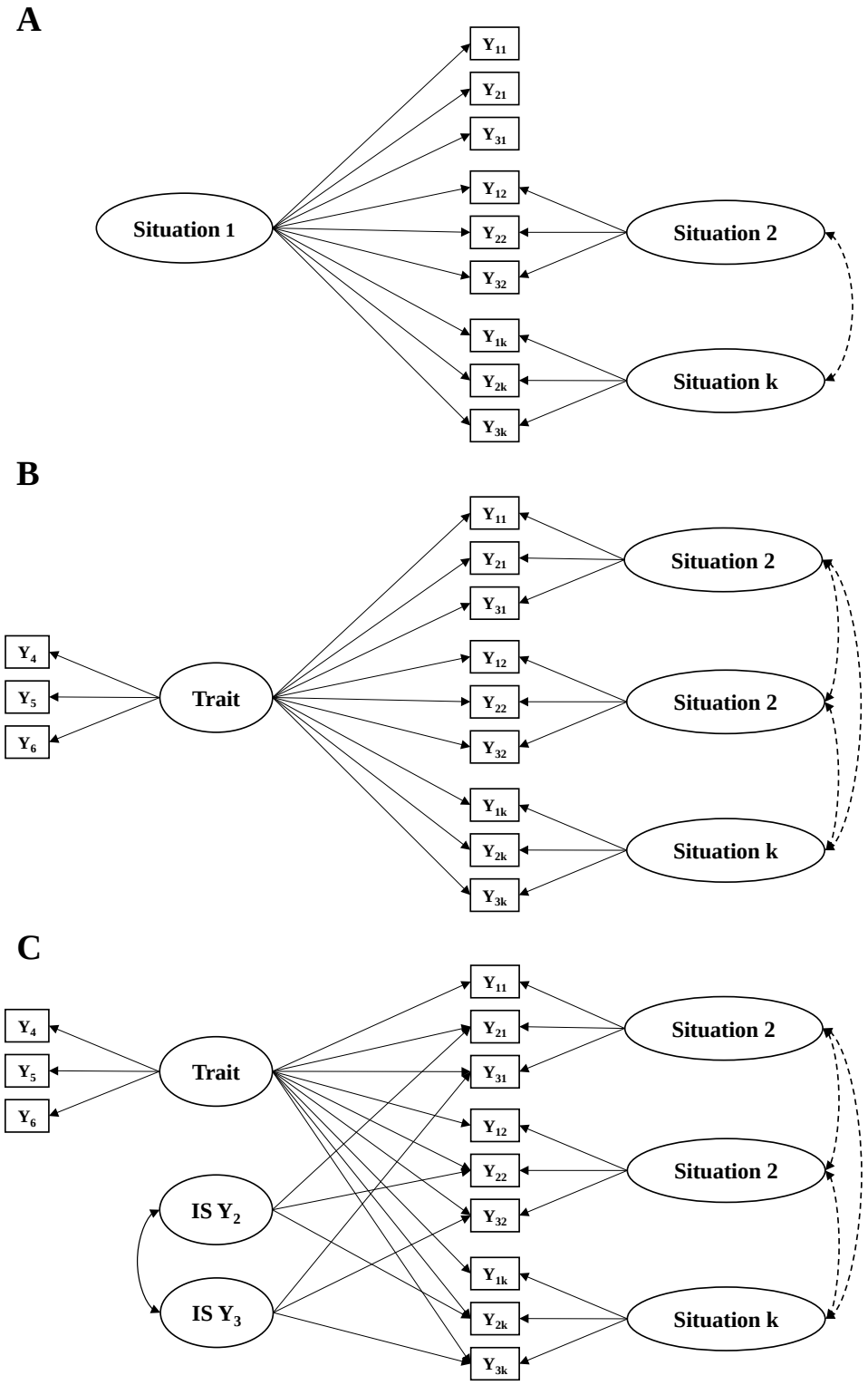
Another way to incorporate a reference is to include a separate trait measure (Eid, 2020). This could be a traditional self-report questionnaire that assesses the corresponding trait to the states measured in the SSA. Alternatively, the state measure of the SSA may be instructed differently to assess general traits. This trait questionnaire reflects the reference for the general trait model on which all items across all situations load. Additionally, situation specific factors for all situations can be specified. Figure 2 (Panel B) reflects such a model for a trait questionnaire and an SSA with three situations. The general trait, as well as all states were assessed with three items each. This model enables to separate trait-specific variance components from responses to SSA situation so that situation-specific factors resemble more closely latent state residuals as described by LST.

However, the interpretation of the latent variables differs slightly from typical LST models. That is, the trait factor reflects the trait questionnaire rather than consistent variance components among state measures. Still, loadings of the trait factor on state indicators represent the consistency of state assessments with the trait assessment. This fine line between LST models and the bifactor S-1 models in the context of SSAs emphasizes the care with which researchers should select the trait assessment methods. That is, trait questionnaires should not only exhibit construct-related validity, but also high symmetry with regard to the situational context of SSA situations (Schulze et al., 2020). For instance, if the SSA consists of a broad variety of situations, a broad trait measure is appropriate. However, if the SSA only comprises a specific class of situations (e.g., work-related situations), the trait measure should reflect the same situational context to maximize the convergence of trait-consistent variance in the SSA and the trait measure. If the SSA captures systematic trait variance that is not shared with the trait measure (e.g., due to validity issues or asymmetry between SSA and trait measure), state-specific residual factors will consist of situation effects, person-situation interaction effects and trait-specific variance that is not shared with the trait questionnaire.

As a consequence, correlated state factors may be needed to achieve acceptable model fit. Such substantial correlations among state residual factors may reflect consistent trait variance that is not captured by the trait measure. Moreover, certain similarities among situations in the SSA may lead to systematic variance components that are shared among state

Figure 2

Measurement Models for Standardized State Assessments



Notes. Exemplary latent variable models for an SSA with three situation descriptions. States within situation descriptions were assessed with three indicators. **A** S-1 model with baseline situation as reference factor. **B** S-1 model with trait measure as reference factor. **C** S-1 model with trait measure as reference factor and indicator-specific method factors. Y describes manifest indicators and k the total number of situations in the SSA. Residual variances and loading parameters are not depicted to avoid clutter.

residual factors but not with the reference trait measure. Such systematic variance components may emerge due to the fixed situations. We suggest to compare a model without state residual factor correlations to a model with correlations. Since these two models are nested, model comparisons are easily implemented via χ^2-difference tests.

Finally, method effects may contribute to systematic variance among state residuals (Eid et al., 1999; Geiser & Lockhart, 2012). That is, due to the repeated assessment of states on identical items, systematic method effects may occur. These effects are not unique to SSA. However, they may be more pronounced in comparison to longitudinal research designs, since time lag between situations in SSA is minimal. The M-1 approach is appropriate to deal with such systematic sources of variance (Eid et al., 1999; Geiser & Lockhart, 2012). Similar to the bifactor S-1 approach, latent indicator-specific factors are defined for all state indicators except one (see Figure 2, Panel C).

Examining the Validity of Standardized State Assessments

The proposed measurement models allow for important inferences about SSA's validity. For instance, if a state measure for a specific situation does not share any substantial variance with the corresponding trait, it may be reasonable to take a closer look at the situation. One reason may be that the situation description failed to address the relevant trait. Another reason could be that responses were mostly driven by the situation and variance between individuals is not sufficient. Furthermore, correlations among latent state residual factors should be closely inspected. That is, between highly divergent situations, no substantial correlation should emerge if trait, method, and current state variance components have been accounted for. However, contextual similarities between situation could explain such correlations (e.g., both situations take place at work).

Beyond the internal structure of SSAs, the most important issue with regard to validity is whether SSAs assess similar states when compared to real-life states. Two general approaches may be feasible to answer this question. First, researchers may mimic SSA situations in a laboratory setting. For instance, if the SSA contains a situation that describes a group discussion, such a group discussion could be simulated in a laboratory setting. Role-players could ensure that all trait-relevant cues of the written SSA situation are transferred to the laboratory setting. Generally, all principles described in the section on situation description should be applied to the laboratory setting (see also Fleeson & Law, 2015).

However, laboratory studies are not equally feasible for all research questions or situations as they are costly, do not allow to simulate all kinds of situations, and lack ecological validity. Hence, researchers may combine SSA with ambulatory assessments. Based on contextual information, real-life situations could be clustered into classes which could be compared to SSA situations that match the same situation class. Generally speaking, the state variability within individuals for a certain context should be similar to the state variability of SSA situations. To test this hypothesis, the LST model for random and fixed situations (Geiser et al., 2015) could be used to disentangle variance components of traits, situations, and person-situation interactions. In this model, randomly sampled situations are nested in fixed situations. The authors use

the example of ambulatory assessments for which contextual information of each situation is available. Sampled situations could then be nested within different contexts (e.g., work-related, home-related, etc.). Similarly, SSA situations and situations from ambulatory assessments may be nested within specific situation classes.

Further relations among states and traits with other constructs and external criteria should be taken into considerations to examine the validity of SSAs. That is, theory-driven nomological nets could shed further light into underlying processes of specific situations. For instance, an agreeableness state for a team situation should correlate more strongly with team performance than a state for a situation with friends or family (see Peeters et al., 2006).

Finally, we suggest experimental test validation to examine the intended effects of situations in SSA (Krumm et al., 2019). As we described above, each situation description should be developed with care and a specific purpose in mind. Thus, each element (or phrase) of a situation description should have an effect on individuals' states. Such effects could be tested by manipulating some aspects of situation descriptions (i.e., changing or deleting critical phrases), while keeping others constant (e.g., apply the same situation descriptions but with different interaction partners). Testing individuals' changes in states between an original and a manipulated situation may give insight on whether states in SSAs are in-fact partly driven by situation effects and person-situation interactions.

Discussion

In this article, we proposed a methodological framework – namely Standardized State Assessment – that uses hypothetical situation descriptions for the assessment of person-situation processes. SSA is designed to provide assessment opportunities beyond the sampling of individuals in real-life situations, as it is typically done in ambulatory assessment studies. Similar approaches have been adopted before (Blum et al., 2018; Kammrath et al., 2005; Lievens, 2017a; Lievens et al., 2018; Rauthmann, 2012; Ziegler et al., 2019). Although all these approaches contributed significantly to a deeper understanding of person-situation processes, overarching guidelines were missing that allow researchers to easily adopt and enhance these assessment methods. Thus, we developed recommendations to develop situation descriptions and state measures to assess person-situation processes. Moreover, we presented latent variable models and validation strategies that are needed to assure that SSAs reflect valuable measures of person-situation processes.

Importantly, contemporary theories about person-situation processes, especially Whole Trait Theory, served as foundation for the SSA framework. That is, we designed SSAs to measure individual distributions of psychological states that reflect dispositional tendencies as well as situation-specific influences. This preposition demonstrates a clear and testable rational about psychological processes that underly SSAs. So, the reliable variance of state measures in SSA should reflect trait-specificity as well as situation-specificity. Further, these components should be related to general tendencies and variability in state measures of real-life ambulatory assessments. Whole Trait Theory also posits that beyond observable state distributions, personality traits

should comprise explanations for these state distributions (Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019). Several theories exist that may be suited to explain trait distributions and that may further enhance an understanding of underlying psychological processes of SSAs or more precise research designs to study these mechanisms (e.g., Mischel & Shoda, 1995). Overall, the theoretical underpinning of the SSA framework enables researchers to draw justified conclusions from SSAs instead of just believing that situation descriptions effectively work as psychological situations.

Even though we postulated that situation descriptions may mimic real-life situations as close as possible, it is not warranted that SSAs actually measure the same processes. The central element of the SSA framework therefore is the theory-driven development of situation descriptions and state measures. We incorporated situation taxonomies into the development of situation descriptions to mimic real-life situations as best as possible. Although similar ideas have been proposed before (Freudenstein et al., 2020; Lievens, 2017a; Lievens et al., 2020), the SSA framework provides specific guidelines on how this may be accomplished. Moreover, the theory-driven development of state measures and use of latent measurement models should enhance the overall quality of SSAs. The use of valid state measures and control of measurement error is especially needed for SSA. That is, beside the typical measurement error of self-report measures, the situation descriptions contribute to a more complex measurement and thus error variance, so that the control of measurement error is inevitable.

A central idea of the SSA framework is that similar psychological processes are being assessed when compared to ambulatory assessments. However, we do not propose that SSA may replace ambulatory assessments. Rather, SSA may be disadvantageous whenever a large number of situations must be sampled and the ecological validity is a primary concern, but may comprise several advantages that may be useful in some contexts. For instance, SSA provides a much more economic approach to sampling persons within situations. Roekel and colleagues (2019) reported an average of 5.65 daily assessments for 12.30 days in ambulatory assessments with adolescents. Those repeated measurements across a day or even weeks increase the burden to participate (Fisher & To, 2012; Roekel et al., 2019; Wright & Zimmermann, 2019), which may be reduced by SSA. Therefore, SSA further facilitates the assessment of person-situation processes in larger studies for which ambulatory assessment is usually not suitable. Similarly, SSA could be applied when researchers are interested in person-situation processes for specific contexts. For example, research on teamwork may be interested in certain personality state dynamics across various team situations. Situation descriptions could then be designed to closely match theory-driven assumptions about these processes. The same is true for research on person-situation processes in rare situations, that were previously unlikely to be sampled, or even completely new situations that so far were not included in individuals' realities but may in the future.

Standardized State Assessment and Situational Judgment Tests

Recently, SJTs have been proposed as alternative method for the assessment of person-situation processes (Lievens, 2017a; Lievens et al., 2020; Martin-

Raugh & Kell, 2019). SJTs typically consist of short situation descriptions with several behavioral response options (Corstjens et al., 2017). Situation descriptions in SJTs are considered essential for individuals' responses (Campion & Ployhart, 2013; McDaniel & Nguyen, 2001; Weekley et al., 2006, 2015). So, SJTs are similar in scope and appearance to the SSA framework. However, for two reasons, we argue that SSA is needed as a separate methodological framework from SJTs to assess person-situation processes.

First, several studies questioned the notion that situational descriptions influence responses to SJT items. For instance, Krumm and colleagues (2015) omitted situation descriptions from SJT items and showed that item difficulty did not depend on the presence of situation descriptions for a large number of items. This was also true for an SJT with video-based situation sequences (Schäpers, Lievens, et al., 2020). Moreover, situation descriptions in SJT items seemed to have negligible effects on the construct-related and criterion-related validity of the measures (Schäpers et al., 2019; Schäpers, Freudenstein, et al., 2020). One explanation to these findings may be that situations in SJT items lack theoretical foundation. Only recently it has been proposed to add trait-activating cues to situation descriptions in order to assess specific traits with SJTs (Guenole et al., 2017; Lievens, 2017b). However, such trait-activating cues did not improve the relevance of situation descriptions for test-takers' responses (Schäpers, Freudenstein, et al., 2020). So, a more severe obstacle to the relevance of situation descriptions in SJT items may lie in the response format. Although exceptions exist, typical SJT items contain response options that describe specific behavioral alternatives for a given situation (Weekley et al., 2015). These response options are in many cases sufficient to construe relevant situations (Freudenstein et al., 2020; Harris et al., 2016; Melchers & Kleinmann, 2016). Even when situation descriptions are available, test-takers often rely on processes based on response options to respond to SJT items (Freudenstein et al., 2020). So, the assessment of person-situation processes is blurred by the availability of situation cues in both situation descriptions and response options (see also Grand, 2019). As delineated above, such specific response options further hinder the comparisons of states between situations as true state change are not distinguishable from method-specific variance. SSA overcomes these problems by providing theory-driven guidelines for the development of situations descriptions, decoupling situation cues from state measures, and standardizing state measures across situations.

Second, SJTs represent a collection of methods rather than a specific assessment of definite constructs (Lievens et al., 2008). Different SJTs vary in their response medium, response format, instruction format, scoring method and more (Campion et al., 2014). The common core of most SJTs is the focus on the prediction of external criteria such as job performance (Christian et al., 2010; Corstjens et al., 2017). A majority of SJTs thereby uses knowledge instructions so that test-takers are asked to report what they *should* do in a given situation rather than what they *would* do (McDaniel et al., 2007). Accordingly, overarching frameworks about psychological processes underlying SJT responses characterized these tests as measures of procedural knowledge (i.e., knowing what to do in specific situations; Lievens & Motowidlo,

2016). However, this conceptualization is in contrast to the assessment of personality states. We believe that, in order to assess momentary states by using hypothetical situation descriptions, a new methodological framework is needed that sets itself apart from the assessment of knowledge and has a theory-driven focus on person-situation processes. SSA may serve as this new methodological framework.

Research Opportunities

Besides practical advantages of SSA, this framework may be useful to tackle several theoretical questions about person-situation processes. First, SSA may be used to scrutinize person-situation contingencies (see also Lievens, 2017a). Since situations in SSA are fixed, researchers are able to uncover direct effects of specific situation cues or cue combinations on state expressions. Previously, such studies only took broad contexts or situation classes instead of specific situations into account (e.g., Fleeson, 2007).

Second, SSA allows for subgroup analyses such as cross-cultural differences or differences among age groups or other demographic variables. The fixed situation design of SSA allows for the analyses of differences between groups in specific situations and thus for more fine-grained interpretations about when and why certain groups differ. For instance, research on intercultural teams may benefit from SSA when studying how different team members perceive specific situations. Similar research designs may be helpful whenever individuals are nested in groups such as families or classes.

Finally, SSA can be applied in longitudinal settings to examine intra-individual processes over time. In comparison to ambulatory assessments, SSA has the advantage that identical situations can be compared over time. For instance, clinical psychologists may be interested in intra-individual processes in various situations before, during, and after episodes of depression in patients. Overall, the SSA framework may be the method of choice whenever the comparison of different responses to identical situations is the focus of research.

Limitations

Although previous research successfully used situation descriptions to study person-situation processes, the SSA framework incorporates several new propositions, which have thus far not been tested and need further research. For instance, SSAs require test-takers to respond to several different hypothetical situations in a short amount of time. Thus, the most demanding tasks for participants is to quickly immerse themselves in different situations. This may increase the likelihood for careless responding. To counteract such unwanted response patterns, additional open response formats may be helpful, which ask participants how they would behave, or what they feel or think in the given situation. Such open response formats demand test-takers to process the given situation descriptions and to think about what they would do in the given situation. Subsequently, test-takers can rate their indicated behavior (or thoughts and feelings) on the state-specific scale (see Runge & Lang, 2019). Overall, this may help to enhance SSA's similarity to real-life state assessments.

The SSA framework builds on the presumption that processes underlying hypothetical situations are similar to real-life situations. Most notably, this assumption is grounded in research by Lievens and colleagues (2018) who demonstrated that

the average and the variability of individuals' responses to an SJT was related to the mean of trait measures and real-life state variability. However, other results may provide reason to act with caution. In particular, the day reconstruction method demonstrated very similar between person variability when compared to ambulatory assessments but differed substantially in regard to within-person variability (Lucas et al., 2019). In the day reconstruction method, participants reconstrue diaries of situations they encountered during the day (Kahneman et al., 2004). Similar to SSA, participants then respond to state measures for each situation. A key difference between the day reconstruction method and SSA is that the former relies on retrospective construction of situations whereas the latter provides theory-driven situation descriptions. Nevertheless, research is needed to examine how SSA relates to state measures in ambulatory assessments.

Another concern may be that situation descriptions in the SSA framework present situations as independent entities. Although this procedure enables much more controlled assessments of person-situation processes, these situations do not reflect reality. Real-life situations are typically defined by constant changes of cues or are shaped by person-situation transactions such as consequences of preceding behavior (see Rauthmann, Sherman, & Funder, 2015). We agree that the SSA framework is not suited to study such complex processes.

We also argued that the SSA framework provides an economic alternative to ambulatory assessment. This is true when considering the data collection phase. However, when following all guidelines proposed in this framework to develop and validate SSAs, the effort to realize SSAs may be much higher compared to ambulatory science, especially when a large number of situation descriptions should be included. Thus, we advocate open science practices and call for researchers to openly share all attempts of SSA developments.

Finally, we concentrated on unidimensional state measures when describing this framework. Multidimensional state measures may generally be possible with the SSA framework. Researchers may have to add various trait-activating cues to assess relevant states. One drawback from this added complexity may be that the control over situations is reduced. If state distributions are the main purpose of the SSA, a multidimensional approach may be fine if several trait-relevant cues were adopted into situation descriptions (see Tett & Guterman, 2000). However, we propose that researchers that are interested in specific situation effects slowly increase the complexity and multidimensionality of situation descriptions in an SSA with the degree to which the mechanism behind certain effects is known and understood.

Conclusion

Many previous approaches used hypothetical situation descriptions to assess person-situation processes. However, these methods often relied on ad-hoc developed situation descriptions and state measures. The SSA framework offers theory-driven guidelines for the development of such methods, including effective ways to assess their validity. We believe that this framework contributes to high quality assessments of person-situation processes whenever ambulatory assessments are unavailable or not suited to examine specific research questions.

References

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Bleidorn, W. (2009). Linking personality states, current social roles and major life goals. *European Journal of Personality*, *23*(6), 509–530. https://doi.org/10.1002/per.731

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*(3), 286–305. https://doi.org/10.1002/per.2138

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures. In Neil D. Christiansen & Robert P. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). Routledge.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, *27*(4), 283–310. https://doi.org/10.1080/08959285.2014.929693

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*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgment tests for selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention* (pp. 228–248). John Wiley & Sons Ltd.

Costello, C. K., Wood, D., & Tov, W. (2018). Revealed traits: A novel method for estimating cross-cultural similarities and differences in personality. *Journal of Cross-Cultural Psychology*, *49*(4), 554–586. https://doi.org/10.1177/0022022118757914

Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor s-1 model for individual clinical assessment. *Journal of Abnormal Child Psychology*. https://doi.org/0.1007/s10802-020-00624-9

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Eid, M., Schneider, C., & Schwenkmezger, P. (1999). Do you feel better or worse? The validity of perceived deviations of mood states from mood traits. *European Journal of Personality*, *13*(4), 283–306. https://doi.org/10.1002/(SICI)1099-0984(199907/08)13:4<283::AID-PER341>3.0.CO;2-0

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*(7), 1097–1126. https://doi.org/10.1037/0022-3514.37.7.1097

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior*, *33*(7), 865–877. https://doi.org/10.1002/job.1803

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, *75*(4), 825–862. https://doi.org/10.1111/j.1467-6494.2007.00458.x

Fleeson, W., & Gallagher, M. P. (2009). The implications of big-five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology*, *97*(6), 1097–1114. https://doi.org/10.1037/a0016786

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009

Fleeson, W., & Law, M. K. (2015). Trait enactments as density distributions: The role of actors, situations, and observers in explaining stability and variability. *Journal of Personality and Social Psychology*, *109*(6), 1090–1104. https://doi.org/10.1037/a0039517

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*(4), 1667–1684. https://doi.org/10.1111/j.1751-9004.2008.00122.x

Freudenstein, J.-P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. https://doi.org/10.1111/peps.12385

Funder, D. C. (2016). Taking situations seriously: The situation construal model and the Riverside Situational Q-Sort. *Current Directions in Psychological Science*, *25*(3), 203–208. https://doi.org/10.1177/0963721416635552

Furr, R. M., & Funder, D. C. (2004). Situational similarity and behavioral consistency: Subjective, objective, variable-centered, and person-centered approaches. *Journal of Research in Personality*, *38*(5), 421–447. https://doi.org/10.1016/j.jrp.2003.10.001

Geiser, C., Hintz, F., Burns, G. L., & Servera, M. (2017). Latent variable modeling of person-situation data. In J. F. Rauthmann, R. A. Sherman, & D. C. Funder (Eds.), *The Oxford handbook of psychological situations*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190263348.013.15

Geiser, C., Litson, K., Bishop, J., Keller, B. T., Burns, G. L., Servera, M., & Shiffman, S. (2015). Analyzing person, situation and person x situation interaction effects: Latent state-trait models for the combination of random and fixed situations. *Psychological Methods*, *20*(2), 165–192. https://doi.org/10.1037/met0000026

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods*, *17*(2), 255–283. https://doi.org/10.1037/a0026977

Grand, J. (2019). A general response process theory for situational judgment tests. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000468

Guenole, N., Chernyshenko, O. S., & Weekly, J. A. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, *17*(3), 234–252. https://doi.org/10.1080/15305058.2017.1297817

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*(1), 23–28. https://doi.org/10.1017/iop.2015.110

Heller, D., Komar, J., & Lee, W. B. (2007). The dynamics of personality states, goals, and well-being. *Personality and Social Psychology Bulletin*, *33*(6), 898–910. https://doi.org/10.1177/0146167207301010

Hintz, F., Geiser, C., & Shiffman, S. (2018). A latent state-trait model for analyzing states, traits, situations, method effects, and their interactions. *Journal of Personality*, *87*(3), 434–454. https://doi.org/10.1111/jopy.12400

Hofmans, J., De Clercq, B., Kuppens, P., Verbeke, L., & Widiger, T. A. (2019). Testing the structure and process of personality using ambulatory assessment data: An overview of within-person and person-specific techniques. *Psychological Assessment*, *31*(4), 432–443. https://doi.org/10.1037/pas0000562

Horstmann, K. T., & Ziegler, M. (2020). Assessing personality states: What to consider when constructing personality state measures. *European Journal of Personality*. https://doi.org/10.1002/per.2266

Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2016). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology*, *90*(1), 1–27. https://doi.org/10.1111/joop.12151

Jayawickreme, E., Zachry, C. E., & Fleeson, W. (2019). Whole trait theory: An integrative approach to examining personality structure and process. *Personality and Individual Differences*, *136*, 2–11. https://doi.org/10.1016/j.paid.2018.06.045

Jones, A. B., Brown, N. A., Serfass, D. G., & Sherman, R. A. (2017). Personality and density distributions of behavior, emotions, and situations. *Journal of Research in Personality*, *69*, 225–236. https://doi.org/10.1016/j.jrp.2016.10.006

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, *306*(5702), 1776–1780. https://doi.org/10.1126/science.1103572

Kammrath, L. K., Mendoza-Denton, R., & Mischel, W. (2005). Incorporating if... Then... Personality signatures in person perception: Beyond the person-situation dichotomy. *Journal of Personality and Social Psychology*, *88*(4), 605–618. https://doi.org/10.1037/0022-3514.88.4.605

Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2018). Explaining general and specific factors in

longitudinal, multimethod, and bifactor models: Some caveats and recommendations. *Psychological Methods*, *23*(3), 505–523. https://doi.org/10.1037/met0000146

Krumm, S., Hüffmeier, J., & Lievens, F. (2019). Experimental test validation: Examining the path from test elements to test performance. *European Journal of Psychological Assessment*, *35*(2), 225–232. https://doi.org/10.1027/1015-5759/a000393

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology*, *100*(2), 399–417. https://doi.org/10.1037/a0037674

Latham, G. P., & Sue-Chan, C. (1999). A meta-analysis of the situational interview: An enumerative review of reasons for its validity. *Canadian Psychology/Psychologie Canadienne*, *40*(1), 56–67. https://doi.org/10.1037/h0086826

Lievens, F. (2017a). Assessing personality-situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*(5), 424–440. https://doi.org/10.1002/per.2111

Lievens, F. (2017b). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing*, *17*(3), 269–276. https://doi.org/10.1080/15305058.2017.1309857

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*(7), 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*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*(5), 1181–1188. https://doi.org/10.1037/0021-9010.91.5.1181

Lievens, F., Schäpers, P., & Herde, C. N. (2020). Situational judgment tests: From low-fidelity simulations to alternative measures of personality and the person-situation interplay. In D. Wood, P. Harms, S. Read, & A. Slaughter (Eds.), *Emerging approaches to measuring and modeling the person and situation*. Elsevier.

Lucas, R. E., Wallsworth, C., Anusic, I., & Donnellan, B. (2019). *A direct comparison of the day reconstruction method and the experience sampling method*. https://doi.org/10.31234/osf.io/cv73u

Martin-Raugh, M. P., & Kell, H. J. (2019). A process model of situational judgment test responding. *Human Resource Management Review*. https://doi.org/10.1016/j.hrmr.2019.100731

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*(1), 63–91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*(4), 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*(1-2), 103–113. https://doi.org/10.1111/1468-2389.00167

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*(1), 29–34. https://doi.org/10.1017/iop.2015.111

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*(1), 121–140. https://doi.org/10.1177/0149206309349309

Mischel, W. (1977). The interaction of person and situation. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333–352). Lawrence Erlbaum.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. https://doi.org/1995-25136-001

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality

dispositions. *Annual Review of Psychology*, *49*(1), 229–258. https://doi.org/10.1146/annurev.psych.49.1.229

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

Mussel, P., Gatzka, T., & Hewig, J. (2018). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment*, *34*(5), 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*(5), 475–476. https://doi.org/10.1002/per.2119

Nestler, S., Geukes, K., & Back, M. D. (2018). Modeling intraindividual variability in three-level multilevel models. *Methodology*, *14*(3), 95–108. https://doi.org/10.1027/1614-2241/a000150

Olaru, G., Burrus, J., Maccann, C., Zaromb, M. F., Wilhelm, O., & Roberts, D. R. (2019). Situational judgment tests as a method for measuring personality: Development and validity evidence for a test of dependability. *PLoS One*, *14*(2), e0211884. https://doi.org/10.1371/journal.pone.0211884

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., de Vries, R. E., & de Wit, M. (2018). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*(1), 1–29. https://doi.org/10.1080/08959285.2018.1539856

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2016). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology*, *112*(4), 642–681. https://doi.org/10.1037/pspp0000111

Peeters, M. A., Van Tuijl, H. F., Rutte, C. G., & Reymen, I. M. (2006). Personality and team performance: A meta-analysis. *European Journal of Personality*, *20*(5), 377–396. https://doi.org/10.1002/per.588

Rauthmann, J. F. (2012). You say the party is dull, I say it is lively: A componential approach to how situations are perceived to disentangle perceiver, situation, and perceiver× situation variance. *Social Psychological and Personality Science*, *3*(5), 519–528. https://doi.org/10.1177/1948550611427609

Rauthmann, J. F. (2015). Structuring situational information. A road map of the multiple pathways to different situational taxonomies. *European Psychologist*, *20*(3), 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., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677–718. https://doi.org/10.1037/a0037250

Rauthmann, J. F., Horstmann, K. T., & Sherman, R. A. (2019). Do self-reported traits and aggregated states capture the same thing? A nomological perspective on trait-state homomorphy. *Social Psychological and Personality Science*, *10*(5), 596–611. https://doi.org/10.1177/1948550618774772

Rauthmann, J. F., Jones, A. B., & Sherman, R. A. (2016). Directionality of person-situation transactions: Are there spillovers among and between situation experiences and personality states? *Personality and Social Psychology Bulletin*, *42*(7), 893–909. https://doi.org/10.1177/0146167216647360

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*(3), 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*(4), 311–329. https://doi.org/10.1177/1088868308321721

Roekel, E. van, Keijsers, L., & Chung, J. M. (2019). A review of current ambulatory assessment studies in adolescent samples and practical recommendations. *Journal of Research on Adolescence*, *29*(3), 560–577.

https://doi.org/10.1111/jora.12471

Runge, J. M., & Lang, J. W. B. (2019). Can people recognize their implicit thoughts? The motive self-categorization test. *Psychological Assessment*, *31*(7), 939–951. https://doi.org/10.1037/pas0000720

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*(3), 479–503. https://doi.org/10.1111/j.1467-6494.2007.00446.x

Schäpers, P., Freudenstein, J.-P., Mussel, P., Lievens, F., & Krumm, S. (2020). Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests. *Journal of Research in Personality*. https://doi.org/10.1016/j.jrp.2020.103963

Schäpers, P., Lievens, F., Freudenstein, J.-P., Hüffmeier, J., König, C. J., & Krumm, S. (2020). 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*, *93*(2), 472–494. https://doi.org/10.1111/joop.12297

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

Schultze, M. (2017). *Constructing subtests using ant colony optimization* [Doctoral dissertation, Freie Universität Berlin]. http://doi.org/10.17169/refubium-622

Schulze, J., West, S. G., Freudenstein, J.-P., Schäpers, P., Mussel, P., Eid, M., & Krumm, S. (2020). *Hidden framings and hidden asymmetries in the measurement of personality – A combined lens-model and frame-of-reference perspective* [Manuscript under review].

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*(5), 872–888. https://doi.org/10.1037/pspp0000036

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Soto, C. J., & John, O. P. (2017). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Stevens, M. J., & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. *Journal of Management*, *20*(2), 503–530. https://doi.org/10.1177/014920639402000210

Steyer, R., Ferring, D., & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*(2), 79–98.

Steyer, R., & Schmitt, M. (1990). The effects of aggregation across and within occasions on consistency, specificity and reliability. *Methodika*, *4*, 58–94.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Ten Berge, M. A., & De Raad, B. (1999). Taxonomies of situations from a trait psychological perspective. A review. *European Journal of Personality*, *13*(5), 337–360. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<337::AID-PER363>3.0.CO;2-F

Ten Berge, M. A., & De Raad, B. (2001). The construction of a joint taxonomy of traits and situations. *European Journal of Personality*, *15*(4), 253–276. https://doi.org/10.1002/per.410

Ten Berge, M. A., & De Raad, B. (2002). The structure of situations from a personality perspective. *European Journal of Personality*, *16*(2), 81–102. https://doi.org/10.1002/per.435

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*(4), 397–423. https://doi.org/10.1006/jrpe.2000.2292

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting

guidelines and current practices. *Journal of Abnormal Psychology*, *129*(1), 56–63. https://doi.org/10.1037/abn0000473

Van Heck, G. L., Perugini, M., Caprara, G.-V., & Fröger, J. (1994). The big five as tendencies in situations. *Personality and Individual Differences*, *16*(5), 715–731. https://doi.org/10.1016/0191-8869(94)90213-5

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 295–322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgment Tests. Theory, measurement and application* (pp. 1–11). Lawrence Erlbaum Associates.

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). Lawrence Erlbaum Associates.

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*(1), 44–63. https://doi.org/10.1080/08959280802540999

Wiedenroth, A., & Leising, D. (2020). What's in an adjective? *Journal of Individual Differences*. https://doi.org/10.1027/1614-0001/a000316

Wilson, R. E., Thompson, R. J., & Vazire, S. (2017). Are fluctuations in personality states more than fluctuations in affect? *Journal of Research in Personality*, *69*, 110–123. https://doi.org/10.1016/j.jrp.2016.06.006

Wright, A. G., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment*, *31*(12), 1467–1480. https://doi.org/10.1037/pas0000685

Ziegler, M. (2014). Stop and state your intentions! *European Journal of Psychological Assessment*, *30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228

Ziegler, M., Horstmann, K. T., & Ziegler, J. (2019). Personality in situations: Going beyond the OCEAN and introducing the Situation Five. *Psychological Assessment*, *31*(4), 567–580. https://doi.org/10.1037/pas0000654

Zimmermann, J., Woods, W. C., Ritter, S., Happel, M., Masuhr, O., Jaeger, U., Spitzer, C., & Wright, A. G. C. (2019). Integrating structure and dynamics in personality assessment: First steps toward the development and validation of a personality dynamics diary. *Psychological Assessment*, *31*(4), 516–531. https://doi.org/10.1037/pas0000625

Chapter 6

General Discussion

This dissertation was concerned with the psychological assessment of person-situation interactions. Since psychological theories increasingly incorporated such processes (e.g., Fleeson & Jayawickreme, 2015; Mischel & Shoda, 1995; Tett & Guterman, 2000), empirical research devoted more efforts to measure person-situation interactions (Hofmans et al., 2019; Trull & Ebner-Priemer, 2014). Along these lines, methods were proposed that measure behavior in naturally occurring real-life situations, but also methods that only require individuals to imagine behavior in hypothetical situations instead of reporting real-life behavior, thoughts, or feelings – such as Situational Judgment Tests (SJTs; Lievens, 2017a). The underlying person-situation interaction processes of SJTs were a particular focus of this dissertation. Due to the good criterion-related validity, SJTs are popular tools for personnel selection (Christian et al., 2010). However, the focus on predictive validity led to a lack of understanding about how SJTs work as selection tools (see Corstjens et al., 2017; Lievens et al., 2008). On the one hand, SJTs are vaguely described as low-fidelity simulations of real-life situations (Motowidlo et al., 1990; Weekley et al., 2015). But beyond the notion of SJTs as simulations, theory-driven person-situation processes were, until recently, completely missing in SJT research (Brown et al., 2016; Campion & Ployhart, 2013; Harris et al., 2016; Schäpers, Mussel, et al., 2019). On the other hand, situation descriptions of SJT items had often negligible effects on SJT responses (Krumm et al., 2015; Schäpers et al., 2020; Schäpers, Mussel, et al., 2019). This led to the assumption that mostly Implicit Trait Policies (ITPs), a context-independent construct, impact SJT response (Krumm et al., 2015; Lievens & Motowidlo, 2016). That is, individuals rely on implicit beliefs about the general effectiveness of trait-related behaviors rather than situation-specific judgments of suitable behaviors (Lievens & Motowidlo, 2016). These opposite positions on underlying psychological processes were widely debated, with various researchers either agreeing to the notion of SJTs as context-dependent measures (e.g., Crook, 2016; Harvey, 2016) or maintaining that what constitutes situations in SJT items has not been adequately considered in SJT research (e.g., Brown et al., 2016; Harris et al., 2016; Melchers & Kleinmann, 2016).

In the introduction to this dissertation, I proposed a working model of SJT responses to summarize the debate about underlying psychological processes. This model first acknowledges that person characteristics, and especially ITPs, may influence SJT responses independently from the specific situation. However, compelling evidence about the conclusiveness of ITPs as construct and its measurement was missing, so that the utility of ITPs for SJT theory remained unclear. Hence, Chapter 2 took a closer look on the construct-related validity of ITPs. The working model of SJT responses further comprises the relevance of situations for SJT responses. Therefore, I built on theory-driven assumptions about real-life person-situation processes to identify falsifiable relations between situations in SJT items and individual responses (e.g., Funder, 2016;

Meyer et al., 2010; Rauthmann et al., 2014; Tett & Guterman, 2000). That is, mostly the perception of situations in SJT items may matter for response behavior (Brown et al., 2016). Individuals may form a construal of situations that is decisive for test-takers' responses. Chapter 3 tested this assumption by scrutinizing whether the situation construal of SJT items predicted response behavior, which test elements of SJT items (i.e., situation descriptions and response options) influenced relevant situation construal, and whether situation construal of SJT items predicted relevant criteria over and above SJT responses. Finally, the working model of SJT responses included assumptions building on trait activation and situational strength theory (Harris et al., 2016; Meyer et al., 2010; Mischel, 1977; Tett & Guterman, 2000). Harris et al. (2016) argued that stronger situations in SJTs would lead to a decrease in the relation of personality traits and SJT responses. Chapter 4 examined this proposition across a large number of SJT items. Overall, Chapters 2 to 4 comprehensively examined the working model of SJT responses and therefore key elements of the debate about SJTs' underlying psychological processes.

Beyond an understanding of underlying processes of SJT responses, this dissertation intended to provide a theory-driven integration of person-situation interactions into standardized psychological testing. To do so, Chapter 5 described Standardized State Assessment (SSA) as a new methodological framework. Building on the notion that hypothetical situation descriptions may evoke similar processes compared to real-life situations, SSA uses situation descriptions to assess person-situation processes. In particular, the SSA framework was designed to measure psychological states.

The remainder of this chapter briefly summarizes the key results of the main focal points of this dissertation, namely ITPs, person-situation processes as underlying SJT responses, and SSA. In addition to study-specific limitations already mentioned in each chapter, I discuss constraining factors to definite conclusion about the three focal points and how future research may help to answer remaining questions. Moreover, I revisit the working model of SJT responses before ultimately reviewing general contributions of this dissertation the future of person-situation driven psychological assessment with situation descriptions for research and practice.

Underlying Processes of Situational Judgment Tests

Implicit Trait Policies and Situational Judgment Tests

ITPs are defined as implicit beliefs about the effectiveness of behaviors due to the trait these behaviors express (Motowidlo et al., 2006b). This construct is conceptually entangled with SJT, as ITPs were originally developed to explain why responses to SJTs correlate with personality traits (Motowidlo et al., 2006a, 2006b). These authors argued that individuals who possess a certain personality trait also tend to belief that behaviors that reflect this trait are generally more effective. When SJT response options reflect a

specific trait, individuals with ITPs for this trait tend to choose these response options when asked about the most effective behavior in a given situation. In fact, several studies demonstrated that measures of ITPs correlate with personality traits (Martin-Raugh et al., 2016; Motowidlo et al., 2006b, 2016, 2018). Building on these initial ideas, Lievens and Motowidlo (2016) reconceptualized SJTs as measures of general domain knowledge. These authors suggested that ITPs reflect general domain knowledge, when the belief about the effectiveness of trait-related behaviors is true for a specific situation (see also Motowidlo & Beier, 2010). As ITPs represent a construct that was developed to explain SJT responses, the assessment of ITPs is also closely related to SJTs. Typically measures of ITPs reflect a different scoring of SJT response that correlates the trait expression of each response option with test-takers' effectiveness ratings (see Lievens, 2017a).

However, this operationalization of ITPs raised several questions about the construct-related validity. Chapter 2 took a closer look at these questions. First, the computational identification of ITPs rather than a direct observation or measurement leads to a potential confound when examining the effects of ITPs on SJT responses. That is, correlations between SJT scores and ITPs, when both scores were derived from the same set of responses, reflect spurious correlations (e.g., Motowidlo & Beier, 2010; Oostrom et al., 2012). A Monte Carlo Simulation demonstrated that this correlation is in fact bound to the saturation of both scoring keys (i.e., to what extent do the SJT response options reflect a certain trait) and thus purely a statistical artifact. Moreover, correlations between SJT and ITP scorers were higher with increasing internal consistencies of either the SJT or ITP score. Overall, these results suggest that the correlation between SJT and ITP scores derived from the same set of responses neither reflects to what degree individuals relied on their ITPs when responding to SJTs, nor to what degree the SJT score is saturated with ITPs. Therefore, the relevance of ITPs for SJT response can only be assessed when ITPs are separately assessed.

The construct-related validity of several measures of ITPs was also assessed in Chapter 2. This was done by testing central theoretical assumptions of ITP measures in a multi-trait multi-method approach (i.e., ITPs for agreeableness and conscientiousness measured with several SJTs). These were the assumption of ITPs as trait-specific, and context-independent constructs, as well as the relation of ITPs and personality traits (see Motowidlo et al., 2006b). However, the results did not strongly support any of these assumptions. That is, most variance in ITP measures was specific to SJTs rather than shared among SJTs, so that the use ITP measures did not generalize across contexts or tests. Further, there was no clear support for a two-factor model of ITPs for agreeableness and conscientiousness. As measures of ITPs were highly test-specific, not enough variance was shared among measures to be crucial for a difference in model fit when only a combined ITP factor was considered. With regard to the relation of ITPs and

personality, Chapter 2 could confirm that personality traits correlated only weakly with measures of ITPs. However, the lack of cohesiveness in measures of ITPs may rather speak to the notion that SJTs correlated with personality traits rather than with the construct of ITPs.

Finally, Chapter 2 reconsidered the notion of ITPs as general domain knowledge. ITPs were defined as general domain knowledge, if trait-specific beliefs about the effectiveness of behavior were correct in a certain situation (e.g., Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010). Building on the epistemological Gettier problem that challenges the notion of justified true beliefs as knowledge (Gettier, 1963), I argued that even though individuals may be justified in their belief that trait-related behavior is effective in a specific situation, the same belief may not hold for other situations. Thus, ITPs may not reflect knowledge. For example, an individuals' ITP for agreeableness may have been developed through general experiences that excluded instances, in which agreeable behavior was ineffective. The second study of Chapter 2 included ITP measures, for which the true effectiveness of response options varied independently from the trait-level of the behavior. The results demonstrated that test-takers rather aligned their responses with the true effectiveness of response options for specific situations than with the trait-level of response options, thus speaking in favor of ITPs as knowledge instead of beliefs. Still, most variance in SJT response could be explained by test-specific factors. Therefore, SJTs may not be pure measures of general domain knowledge in the form of ITPs.

Overall, Chapter 2 demonstrated that measures of ITPs severely lack construct-related validity. As theory and measurement of ITPs have been exclusively associated with SJTs, the lack of construct-related validity in measures of ITPs questions the general relevance of ITPs for SJTs itself. That is, results of previous studies investigating the relation of ITPs and SJTs may rather provide information about specific SJTs than the broader concept of ITPs. For instance, Motowidlo and colleagues (2006b) found that their measures of ITPs correlated with personality traits. However, each ITP was only assessed with a single SJT. Against the background of the results presented in this dissertation, it is questionable if their SJT did in fact measure ITPs. Indeed, most studies on ITPs used a specific SJT to assess ITPs. So, based on the assumption that SJTs are no valid measures of ITPs, compelling evidence for ITPs as underlying construct of SJT responses is lacking. Even the reconceptualization of SJTs as measures of ITPs widely argues based on indirect findings such as the irrelevance of situation descriptions for SJT response (Lievens & Motowidlo, 2016). In sum, this reconceptualization may have been unjustified.

Limitations and Further Research

A main limitation to the generalizability of results presented in Chapter 2 may be the way ITPs were assessed. Previous research argued that all SJTs could be scored to

assess ITPs, if the trait-level of each response option is known (Lievens, 2017a; Lievens & Motowidlo, 2016). Almost all SJTs used in Chapter 2 were designed to serve some other purpose than the assessment of ITPs (e.g., measuring personality traits or teamwork knowledge). Only one SJT was specifically developed to measure ITPs for agreeableness. Hence, one may argue that the lack of construct-related validity in measures of ITPs in Chapter 2 could be partly due to the fact the test developers did not consider ITPs as relevant construct when designing those SJTs. However, no other method to assess ITPs has been proposed. In fact, the assumptions that SJTs may be scored to assess ITPs purely relies on conceptual arguments (e.g., Lievens, 2017a). This is even true for SJTs that were specifically developed to assess ITPs (e.g., Motowidlo et al., 2006b, 2016). Given this lack of empirical evidence on the construct-related validity of ITP measures, Chapter 2 even questions the existence of ITPs as a construct. So, valid measures are indispensable before drawing conclusions about the concept of ITPs.

One merit of future research could be to reconsider the definition and conceptualization of ITPs. That is, ITPs should be detached from SJTs. Researchers should build on real-life assumptions and concepts that contribute to a clear and verifiable definition of ITPs. This definition should especially clarify whether ITPs reflect a general belief of knowledge about the effectiveness of behaviors. As starting point, also other constructs such as the likability of personality traits could be helpful (see Lamkin et al., 2018). Such a definition of ITPs as broad and global construct would overcome the conceptualization as post-hoc explanation of SJT responses. It would further simplify the development of new assessment methods independent from SJTs. Depending on the nature of the revised definition of ITPs, such measure may adopt methods for the assessment of knowledge (e.g., Schipolowski et al., 2013) or implicit and explicit beliefs (e.g., Butler et al., 2007; Nosek et al., 2005).

However, some readers may also argue that concluding that challenges the existence of ITPs may not be warranted given the results presented in Chapter 2. Although a large proportion of variance was specific to SJTs, some variance was shared among SJTs, which could be interpreted as ITP variance. Motowidlo and colleagues (2018) demonstrated similar results for several measures of prosocial ITPs and did in fact derive at a different conclusion. These authors argued that prosocial ITPs were the underlying construct of several SJTs with interpersonal context. Hence, future research should consider alternative explanations why different SJTs correlate with each other. For example, similar contexts may contribute to similar response behavior of test-takers among different SJTs or general mental ability may serve as common core of SJT scores (see McDaniel et al., 2007). Additionally, ITPs were developed to explain why SJTs correlate with personality traits. As alternative explanation to ITPs, one may argue that even when individuals are asked what they *should* do in a given situation (instead of what they *would* do), they may rely on behavioral dispositions to respond to SJT items. Previous

research showed that would-do and should-do instructions showed small to medium correlations (Ployhart & Ehrhart, 2003). Such alternative explanations may be directly tested to further investigate the relation of ITPs and SJTs.

Person-Situation Processes in Situational Judgment Tests

Previous research on person-situation processes in SJTs predominantly focused on situation cues (e.g., Krumm et al., 2015; Schäpers et al., 2020; Schäpers, Mussel, et al., 2019; see also Brown et al., 2016). For example, Schäpers, Mussel and colleagues (2019) recently manipulated the availability of situation cues in SJT items (i.e., situation descriptions) and concluded that the lack of differences in construct-related correlations demonstrates the negligible role of situation construal for SJT responses. The situation construal model (Funder, 2016) is closely related to recent research that disentangles situational information (e.g., Rauthmann, 2015). That is, the objective situations consists of several situation cues whereas situation characteristics describe an individual's perception of a situation (Rauthmann, 2015). Building on this research, Chapter 3 contained three studies that were the first attempt to directly assess situation construal of SJT items. These studies included a large number of SJT items from various construct domains (i.e., knowledge, applied social skills, and personality; see Christian et al., 2010). For each SJT item, participants were asked to rate their situation perception on a standardized measure of situation characteristics. All three studies supported the relevance of situation construal for SJT responses and thus the notion of person-situation interactions as underlying psychological processes of SJTs (see Brown et al., 2016; Martin-Raugh & Kell, 2019). So far, only one other study directly related person-situation processes to SJT responses. Lievens and colleagues (2018) demonstrated that intraindividual variability of responses across SJT items correlated with intraindividual variability in real-life state assessments across several situations, which also supports the relevance of person-situation processes for SJT responses. This consistent evidence across studies that directly assessed person-situation processes supports a situation dependent perspective on SJT.

However, Chapter 3 also revealed some caveats regarding the generalizability of the notion that SJTs measure person-situation processes. For one, the relevance of situation construal varied across SJT items with strong effects for some items whereas other items did not reflect such person-situation processes. Moreover, Chapter 3 demonstrated that the relevance of situation construal for SJT responses was not contingent on the availability of situation descriptions. In a between-subjects design, the relevance of situation construal for SJT responses was compared between groups that saw complete SJT items, only response options, or only situation descriptions. Surprisingly, situation cues in response options were sufficient to construe psychologically relevant situations for most items, regardless of whether previous research demonstrated that situation

descriptions significantly affected those items' difficulty (see Krumm et al., 2015; Schäpers, Mussel, et al., 2019). In other words, the underlying psychological processes remained the same although situation descriptions were omitted. Surprisingly, perceived situation characteristics did not predict SJT responses, when the situation construal was based only on situation descriptions. This result further stresses the relevance of response options for situational processes in SJTs. Finally, results of Chapter 3 showed that situation construal of SJT items predicted relevant criteria over and above SJT responses. That is, situation construal of SJT items contained valuable variance components that were not captured in responses to SJT items. In sum, Chapter 3 speaks to the notion of person-situation interactions as underlying processes of SJT responses, but more research is needed to identify possible moderators that explain when and why situation construal predicts SJT responses.

Importantly, these results may help to bridge the gap between research with seemingly opposing results; especially research that reframed SJTs as context-independent measures. Most notably, several studies demonstrated that situation descriptions in SJTs are often less relevant than previously assumed and concluded that some other processes than person-situation interactions must take place (Kaminski et al., 2019; Krumm et al., 2015; Schäpers et al., 2020; Schäpers, Mussel, et al., 2019). Against the background of Chapter 3, it seems likely that person-situation processes do take place but situation descriptions are not always needed to evoke these processes. That is, response options of SJT items contain sufficient situation cues (see also Harris et al., 2016; Melchers & Kleinmann, 2016). Similarly, Rockstuhl and colleagues (2015) argued that person-situation processes (dubbed as situational judgment) was not captured by SJT responses. Yet, a more fine-grained conclusion may be needed, as SJT responses do capture some person-situation processes but relevant processes remain hidden if not directly assessed (e.g., in the form of situation characteristics).

Beyond situation construal, this dissertation also considered person-situation processes as postulated in Trait Activation Theory (Tett & Guterman, 2000) as underlying mechanisms of SJT responses. Trait Activation Theory posits that personality-relevant behavior can only be observed if situations contain trait activating cues (Tett & Guterman, 2000). It further encompasses the notion of situation strength by arguing that individual differences in personality traits are less relevant for behavior if the situation conveys clear expectations about adequate behavior (Mischel, 1977; Tett & Guterman, 2000). Chapter 4 tested across a large number of SJT items whether situational strength of SJT items moderated the relation of personality and SJT responses. However, the results did not support the notion that personality traits were less relevant for response behavior in stronger SJT situations. Moreover, Appendix C testes in a within-subjects design whether trait activating cues in situation descriptions influenced the construct-related validity of an SJT measuring narrow personality facets. To do so, SJT versions

with and without situation descriptions were compared in their relation to self and peer-reported personality. Results showed that the two test versions did not differ in their relation to other personality measures, so that the study did not support an enhancement of construct-related validity due to trait-activating cues in situation descriptions. Therefore, these results are in contrast to assumptions of Trait Activation Theory and do not support the idea of person-situation interactions as underlying processes of SJT responses.

Limitations and Further Research

As delineated above, this dissertation provided evidence in favor of conceptualizing SJTs as measures of person-situation processes. However, this conclusion was not supported across all studies. So, several limitations should be addressed that may help to further understand these mixed findings. First, the study presented in Appendix C only considered trait-activating cues in situation descriptions to examine whether these cues enhance the construct-related validity of an SJT. This procedure does not account for trait-activating cues in response options (see Harris et al., 2016; Melchers & Kleinmann, 2016). In fact, Chapter 3 demonstrated that person-situation processes may take place even if situation descriptions were omitted. Another recent study took this into account and controlled for the availability of trait-activating cues in situation descriptions and response options (Schäpers, Lievens, et al., 2019). The results showed that trait-related responses to SJT items were more likely when test-takers saw the trait-activating cues. An important implication from these results is that future research on underlying processes of SJT responses should consider the relevance of single situation cues rather than broad test elements (i.e., situation descriptions and response options).

Second, Chapter 4 assumed that some objectivity to situational strength of SJT items exists, which may influence the relation of personality and SJT responses. Thus, situational strength was operationalized as the average rating of subject matter experts. However, situational strength may rather reflect some sort of situation construal in that only the perception of situational strength is relevant as moderator of the relation between personality and SJT responses (see Meyer et al., 2014). Future research should assess test-takers' perception of situational strength. Relatedly, situation construal in Chapter 3 was measured with the DIAMONDS taxonomy (Rauthmann et al., 2014). Although this taxonomy was developed as broad framework of situation characteristics, which should applicable to a broad range of situations (Rauthmann et al., 2014), situations in SJT items may differ to real-life situations. For example, situation descriptions in SJT items are often very broad and unspecific, especially when compared to real-life situations. Hence, some facets of situation taxonomies, such as Mating ("Potential romantic partners are present"; Rauthmann & Sherman, 2016b, p. 166) may be odd in certain situations. Different situation taxonomies may be more suitable for hypothetical situations. For example, the CAPTION taxonomy included the typicality of situations

that encompasses the "commonness and straightforward nature of the situation" (Parrigon et al., 2016, p. 657). Importantly, these authors suggested that the typicality of situations may be relevant to explain state deviations from an individual's trait disposition.

Third, the response format of SJT items may have contributed to mixed findings on the relevance of person-situation processes for SJT responses. For example, most of the SJTs used in this dissertation instructed test-takers to pick a single response option. However, such single responses do not allow to estimate the amount of measurement error for specific situations. If responses to SJT items are substantially driven by measurement error, the relation between situation construal and SJT responses may be attenuated. In fact, the development of a short-form SJT presented in Appendix A provided initial evidence that responses to SJT items are prone to measurement error. In this SJT, test-takers were asked to rate the effectiveness of all response options. All response options were designed to reflect either agreeable or disagreeable behaviors. However, internal consistencies for ratings within situation descriptions were quite low with an average of $\alpha = 38$. Future research should adopt response formats that allow to take measurement error into account. Such response formats also have the tendency to result in higher construct-related validity of SJT scores (Olaru, Jankowsky, et al., 2019).

Finally, studies on underlying processes of SJT responses presented in this dissertation did not control for response instructions of SJTs. That is, typical SJTs either ask for behavioral tendencies (i.e., "What would you do?") or knowledge (i.e., "What should you do?"; McDaniel et al., 2007). The majority of SJT items used in this dissertation asked for behavioral tendencies. This instruction should more closely resemble real-life person-situation processes in contrast to a knowledge instruction (McDaniel et al., 2007). Surprisingly, situation construal also predicted responses to some items used in Chapter 3 that were applied with a knowledge instruction. This may reflect a potential overlap of underlying processes between the two instructions (Ployhart & Ehrhart, 2003). Further research is needed that examines differences in the relevance of person-situation processes between SJT items with different response instructions.

In sum, a follow-up study is needed that takes these limitations into account. SJT items used in this study should be carefully redesigned so that trait-activating cues are only present in situation descriptions and not in response options (see Schäpers, Lievens, et al., 2019). Furthermore, participants should be required to rate all response options with regard to how likely they would engage in the described behavior in the specific situations. To do so, pretests are needed that ensure that response options for a specific situation description reflect a unidimensional construct with sufficient internal consistency. To compare behavioral tendency instructions with knowledge instructions, a within-subjects design may be useful in which participants also rate response options of the same SJT items with regard to their effectiveness in the given situation. Finally, various measures of situation construal should be applied for each SJT item (e.g., Meyer et

al., 2014; Parrigon et al., 2016; Rauthmann & Sherman, 2016a). As such a study would be very demanding and time-consuming for participants, planned missingness designs may be useful to reduce the burden to participate (Graham et al., 2006; Rhemtulla & Hancock, 2016). For example, participants may only respond to some SJT items or some measures of situation construal.

Reassessing the Working Model of Situational Judgment Test Responses

The working model of SJT responses proposed in the introduction to this dissertation summarized the various theoretical arguments and empirical evidence about underlying psychological processes of SJTs (e.g., Harris et al., 2016; Krumm et al., 2015; Lievens & Motowidlo, 2016; Motowidlo et al., 1990). In sum, this dissertation found support for the notion that person-situation processes determine SJT responses and that context-independent processes may not be as relevant as previously thought. Especially Lievens and Motowidlo (2016) argued in favor of the context-independency of SJTs when they reconceptualized SJTs as measures of ITPs. However, Chapter 2 provided compelling evidence that questions the construct-related validity of ITPs and ultimately the concept of ITPs itself. Moreover, Lievens and Motowidlo (2016) based their arguments on empirical studies that demonstrated that SJTs did not work as intended (Krumm et al., 2015; Rockstuhl et al., 2015). Particularly, the irrelevance of situation descriptions for a large amount of SJT items was interpreted in favor of a context-independent perspective of SJTs (Krumm et al., 2015; Lievens & Motowidlo, 2016; Schäpers, Mussel, et al., 2019). Nevertheless, this dissertation provided evidence that integrates these results without discarding person-situation interactions as underlying processes of SJT responses. That is, response options were shown to be much more relevant to situational processes than previously assumed (see Harris et al., 2016; Melchers & Kleinmann, 2016). Overall, I argue that context-independent processes should be excluded from the working model of SJT responses. That is, most research investigating ITPs, the proposed context-independent constructs, is subject to a lack of construct-related validity of applied measures (cf. Motowidlo et al., 2018). Moreover, indirect and conceptual arguments in favor of a context-independent perspective (see Lievens & Motowidlo, 2016) may be integrated into a situation-dependent perspective of SJTs. Thus, future research should focus on person-situation processes. Although this conclusion is in line with the initial notion of SJTs as low-fidelity simulations (Motowidlo et al., 1990), SJTs must overcome this simplification and integrate theory-driven processes.

Having said that, results of this dissertation also revealed several inconsistencies with regard to person-situation interactions as underlying processes of SJT responses. Above I discussed several limitations that may have contributed to these findings. On the other hand, different processes that go beyond person-situation interactions may

determine SJT responses. Other scholars proposed process models to encompass such complex psychological processes (Grand, 2019; Ployhart, 2006). For example, Grand (2019) posited that the response process to SJT items is divided into several sequential parts. First, test-takers interpret the demands of situation descriptions presented in SJT items. Second, individuals decide what they would do in this situation and what consequences they expect from this behavior. Third, individuals judge the expected consequences of response options presented in the SJT item. Finally, participants compare the expected consequences for their own behavior with those of the response options. If this evaluation exceeds a certain similarity threshold, individuals pick the corresponding response option. Especially this last component of the process model is noteworthy. Whereas this comparison of expected consequences for the desired behavior and behaviors described in response options could explain underlying processes beyond person-situation interactions, it lacks justification why such processes are advantageous to assess in the first place. So, instead of trying to understand how individuals respond to SJT items, a different approach is needed that develops psychological assessments based on predefined theoretical assumptions.

Standardized State Assessment

The SSA framework described in Chapter 5 was designed to provide such a theory-driven assessment tool. It builds on Whole Trait Theory that describes personality as a density distribution of psychological states (Fleeson & Jayawickreme, 2015). So, similar to real-life state assessments (i.e., ambulatory assessments), SSA serves the purpose to measure psychological states based on descriptions of hypothetical situations. These situation descriptions are similar to those in SJTs. However, SSA goes beyond the method of SJTs. First, the framework adopts recent research on situations to provide guidelines on the development of situation descriptions. Most notably, situation taxonomies are used to determine what constitutes adequate situation descriptions. Second, behavioral response options were eliminated to standardize state measures across situations and therefore increasing comparability among situations. Instead, adjective markers are used to assess situation-specific states. In comparison to SJTs, these state measures should additionally reduce the situational information in response options and thus increase the relevance of situation descriptions for test-takers responses. Third, specific latent variable models were proposed referring to Latent State Trait Theory (Steyer et al., 1992).

Compared to SJTs, the main advantage of the SSA framework is that it is based on specific theoretical assumptions. For example, in-line with Whole Trait Theory, substantial intraindividual variability in states should exist that should correlate with real-life state variability. These assumptions are falsifiable and thus enable researchers to

decide whether each situation description is well suited to assess the intended psychological processes. Beyond the theory-driven approach of SSA, the framework further draws the attention back to situation descriptions, as situation cues are omitted from response options.

SSA may be a viable alternative for research on person-situation processes, especially when more complex and expensive approaches are not feasible (i.e., ambulatory assessment). In particular, researchers may use SSA to investigate person-situation contingencies (see also Lievens, 2017a), as the theory-driven development of situation descriptions allows for manipulations of selective and meaningful situation cues. Furthermore, SSA facilitates the comparison of situation-specific states between groups or across measurement occasions. However, empirical evidence about the utility of the SSA framework is needed.

Limitations and Further Research

Although the SSA framework builds on promising research that used situation descriptions to assess person-situation processes (e.g., Lievens et al., 2018; Rauthmann, 2012; and studies presented in this dissertation), it currently remains untested and empirical research is needed that develops SSAs and examines their validity. When doing so, an additional aspect should be considered, beyond the procedures to assess the validity of SSAs' discussed in Chapter 5. Whole Trait Theory posits that personality traits are best described as density distributions of states (Fleeson & Jayawickreme, 2015). The mean of this distribution thereby reflects an individual's behavioral tendency, commonly conceptualized as trait (Fleeson & Jayawickreme, 2015). However, this behavioral tendency has not yet been incorporated into the SSA framework. Following Latent State Trait Theory (Steyer et al., 1999), the focus was to rather distinguish trait-relevant variance in state measures from latent state residual variance. Nevertheless, composite scores may be computed for state measures in an SSA. This composite score should reflect general behavioral tendencies and should correlate with aggregate scores of states in ambulatory assessments as well as trait self-reports.

One drawback of the SSA framework is that situations do not naturally occur but have to be carefully designed on the basis of situation cues. In recent years, psychological research on situations focused on individuals' perceptions of situations (i.e., situation characteristics). Situation characteristics have the advantage that different situations are comparable among each other without relying on specific situation cues (Rauthmann et al., 2015). However, to develop situation descriptions that mimic real-life situations, an understanding about what cues contribute to psychologically relevant perceptions of situations is needed. Taxonomies of situation cues (e.g., Saucier et al., 2007) provide first steps towards such an understanding, but the link between situation cues and characteristics often remains hidden. For example, Rauthmann et al. (2014) only found

relatively small correlations between the presence of certain situation cues and perceived situation characteristics. Thereby, the clearest link emerged for work-related cues and the perception of Duty ("Work has to be done"). Similarly, Chapter 3 failed to identify situation cues in SJT item that contributed to the perception of specific situation characteristics, except for the facet Duty. Hence, more research is needed that examines what situation cues are needed to develop situation descriptions that closely resemble real-life situations. For instance, participants of an ambulatory assessment could be required to provide detailed descriptions of situations they currently encounter. Based on these descriptions, an SSA could be developed. Researchers may then compare state assessments between the real-life situation and the SSA. Two factors may be manipulated to examine how situation descriptions are best developed. First, the availability of certain situation cues in the SSA may contribute to the amount of convergence between the two state measures, and second, the specificity of cues or the richness of details may be manipulated.

Relatedly, more research is needed to understand how people may immerse into situations of SSAs. That is, some test-taker may try to imagine very specific situations while others try to think about similar situations to the one described and respond according to the aggregate of states in those similar situations. Think-aloud studies may help to clarify how individuals' approach SSAs (see Krumm et al., 2015). In a next step, such think-aloud studies could be combined with experimental manipulations of test properties. For example, test-takers may more likely imagine a specific situation if they are instructed not to think about similar or past situations. In addition, highly specific situation descriptions may increase the difficulty for test-takers to think about similar situations, which may increase the level of immersion.

Beyond research questions regarding the validity of SSA, this framework may also be useful to gather knowledge about psychological situations. For example, when situations are defined as individuals' perception of situations, the measurement of situations is confounded with the person and thus subject to circularity (Rauthmann et al., 2015). To overcome this impairment, the objective situations may be defined as the agreement between several perceptions of the same situations (Rauthmann et al., 2014, 2015). However, such ratings often rely on open-ended descriptions of situations from participants (e.g., Rauthmann et al., 2014). The SSA framework allows for a much more control over several ratings of situation characteristics for the same situations. By manipulating situation descriptions, it also enables research designs that examine what factors contribute to higher agreement among raters in their perception of situations.

Situational Judgment Tests are Dead, Long Live Situational Judgment Tests

In this dissertation, I proposed a new framework to assess psychological states (i.e., SSA). As this framework contains clear suggestions for the development of situation descriptions and postulates testable assumptions regarding its validity, one may argue that SJTs are no longer needed. However, I propose that both methodological frameworks may coexist as they may serve different purposes. Whenever researchers or practitioners aim at assessing personality states with the help of situation descriptions, an SSA should be developed. This implicates that SJTs no longer adopt behavioral tendency instructions, as these instructions intend the assessment of situation-specific states. This also includes the emerging field of construct-driven SJTs (see Guenole et al., 2017; Lievens, 2017b) of personality traits (e.g., Mussel et al., 2018; Olaru, Burrus, et al., 2019; Oostrom et al., 2018). These tests may be revised to reflect SSAs so that they enable much more control over situation-specific states and, in line with Whole Trait Theory, theory driven assessments of traits.

Nevertheless, SJTs may be useful as measures of job knowledge to predict job performance. That is, most SJTs use some sort of knowledge instruction (e.g., "What would you do?") to assess relevant job knowledge (McDaniel et al., 2007). In particular, these SJTs are designed to measure procedural knowledge "about how to behave in a way that impacts the context in which work gets done" (Torres & Beier, 2016, p. 52). As the construct-related validity of these tests has been described as a "hot mess" (McDaniel et al., 2016, p. 48), I propose several improvements to SJTs as selection tools. First, similar to SSA, systematic attention to specific situations is needed. Contrary to previous SJT developments, the goal should be to assess job knowledge validly and reliably within a series of situations and not across situations. To do so, response formats are needed that capture test-takers' responses on several indicators. For example, individuals may be asked to rate the effectiveness of all response options. Importantly, this procedure may require a larger number of response options during the development stage of an SJT to be able to select a valid and unidimensional set of response options. Only if this condition is met, researchers may investigate the dimensionality of job knowledge across situations. However, as SJTs have been previously described as multidimensional measures (e.g., Lievens, 2017b), it is likely that different relevant situations require different facets or domains of procedural job knowledge. Hence, current SJT scores may be better understood as composite scores of different situational measures that predict job performance. Researchers may even consider weighted composite scores with regard to the individual contribution of specific situations to the prediction of relevant criteria. Second, situation descriptions in SJTs were often demonstrated to be irrelevant, as sufficient contextual information is provided in the response options (e.g., Krumm et al., 2015). Hence, a risk with SJTs as measures of job

knowledge is that test-takers rate the effectiveness of response options without taking the specific situation into account. One remedy may be that SJT items only include response options that are in general effective but only one is effective for the specific SJT item.

Overall, these suggestions are only valuable for practitioners in the medium term as further research is needed to examine whether these revised SJTs are valid predictors of job performance. In the meantime, I urge practitioners to rely on construct-driven SJTs (Guenole et al., 2017; Lievens, 2017b). Although I argued that these tests should be modified into SSAs, construct-driven SJTs exist that validly assess personality traits (e.g., Mussel et al., 2018; Oostrom et al., 2018). Further research is needed to examine the underlying processes on the item level, but psychological processes that contribute to the overall test score are relatively well understood. Practitioners may use these SJTs as contextualized predictors of job-performance, similar to the frame-of-reference technique (Shaffer & Postlethwaite, 2012).

Conclusion

In recent years, the underlying processes of SJTs were subject to debate. While an increasing number of studies suggested, against the notion of SJTs as low-fidelity simulations, that these tests reflect context-independent measures (Jackson et al., 2016; Krumm et al., 2015; Lievens & Motowidlo, 2016; Schäpers, Mussel, et al., 2019), others maintained that situational processes are relevant for SJTs (e.g., Brown et al., 2016; Harris et al., 2016). This dissertation took a much-needed closer look at underlying psychological processes of SJTs. Importantly, I considered SJTs in the context of real-life person situation processes and other assessment methods that aim to measure these processes. Overall, the results were in favor of person-situation interactions as underlying processes of SJTs and provided evidence against the relevance of proposed context-independent constructs (i.e., ITPs). Especially the situation construal, an individual's perception of situations, helped understanding how test-takers process SJT items. Furthermore, the results suggested that recent evidence that demonstrated negligible effects of situation descriptions for SJT responses could be integrated into a perspective of SJTs as measures of person-situation processes. That is, test-takers often rely on response options to construe psychologically relevant situations. However, this dissertation also uncovered major inconsistencies in the notion of person-situation processes as underlying processes of SJTs. For example, key assumptions of Trait Activation Theory were not supported in two studies. Several characteristics of SJTs, such as the nature of response options, or response instructions and formats, may have contributed to these mixed findings. So, the major conclusion of this dissertation is that SJTs are too broad and general as a methodological framework to integrate a common understanding of

underling psychological processes. Rather, new and theory-driven frameworks are needed that build on SJTs but are designed to assess predefined and clear psychological processes. As one such methodological framework, I proposed SSA for the assessment of psychological states based on situation descriptions. Other framework may concentrate on the assessment of procedural job knowledge. Future research must now examine whether these frameworks do in fact assess their proposed underlying processes. In sum, such narrow frameworks may contribute to theory-driven and valid assessments of person-situation processes based on situation descriptions. Thereby, these frameworks may sustain similar advantages than SJTs like economic applications or control over sampled situations.

References

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*(1), 38–42. https://doi.org/10.1017/iop.2015.113

Butler, A. C., Beck, A. T., & Cohen, L. H. (2007). The personality belief questionnaire-short form: Development and preliminary findings. *Cognitive Therapy and Research*, *31*(3), 357–370. https://doi.org/10.1007/s10608-006-9041-x

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures. In Neil D. Christiansen & Robert P. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). Routledge.

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*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgment tests for selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention* (pp. 228–248). John Wiley & Sons Ltd.

Crook, A. E. (2016). Unintended consequences: Narrowing SJT usage and losing credibility with applicants. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *9*(1), 59–63. https://doi.org/10.1017/iop.2015.118

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009

Funder, D. C. (2016). Taking situations seriously: The situation construal model and the Riverside Situational Q-Sort. *Current Directions in Psychological Science*, *25*(3), 203–208. https://doi.org/10.1177/0963721416635552

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323. https://doi.org/10.1037/1082-989X.11.4.323

Grand, J. (2019). A general response process theory for situational judgment tests. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000468

Guenole, N., Chernyshenko, O. S., & Weekly, J. A. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, *17*(3), 234–252.

https://doi.org/10.1080/15305058.2017.1297817

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*(1), 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*(1), 63–71. https://doi.org/10.1017/iop.2015.119

Hofmans, J., De Clercq, B., Kuppens, P., Verbeke, L., & Widiger, T. A. (2019). Testing the structure and process of personality using ambulatory assessment data: An overview of within-person and person-specific techniques. *Psychological Assessment*, *31*(4), 432–443. https://doi.org/10.1037/pas0000562

Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2016). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology*, *90*(1), 1–27. https://doi.org/10.1111/joop.12151

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*, *27*(1), 72–82. https://doi.org/10.1111/ijsa.12233

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*(2), 399–417. https://doi.org/10.1037/a0037674

Lamkin, J., Maples-Keller, J. L., & Miller, J. D. (2018). How likable are personality disorder and general personality traits to those who possess them? *Journal of Personality*, *86*(2), 173–185. https://doi.org/10.1111/jopy.12302

Lievens, F. (2017a). Assessing personality-situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*(5), 424–440. https://doi.org/10.1002/per.2111

Lievens, F. (2017b). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing*, *17*(3), 269–276. https://doi.org/10.1080/15305058.2017.1309857

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*(7), 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*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Martin-Raugh, M. P., & Kell, H. J. (2019). A process model of situational judgment test responding. *Human Resource Management Review*. https://doi.org/10.1016/j.hrmr.2019.100731

Martin-Raugh, M. P., Kell, H. J., & Motowidlo, S. J. (2016). Prosocial knowledge mediates effects of agreeableness and emotional intelligence on prosocial behavior. *Personality and Individual Differences*, *90*, 41–49. https://doi.org/10.1016/j.paid.2015.10.024

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*(1), 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: Perspectives on Science and Practice*, *9*(1), 47–51. https://doi.org/10.1017/iop.2015.115

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*(1), 29–34.

https://doi.org/10.1017/iop.2015.111

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*(1), 121–140. https://doi.org/10.1177/0149206309349309

Meyer, R. D., Dalal, R. S., José, I. J., Hermida, R., Chen, T. R., Vega, R. P., Brooks, C. K., & Khare, V. P. (2014). Measuring job-related situational strength and assessing its interactive effects with personality on voluntary work behavior. *Journal of Management*, *40*(4), 1010–1041. https://doi.org/10.1177/0149206311425613

Mischel, W. (1977). The interaction of person and situation. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333–352). Lawrence Erlbaum.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. 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*(2), 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*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance*, *29*(4), 331–346. https://doi.org/10.1080/08959285.2016.1165227

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). 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). Lawrence Erlbaum Associates.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*(4), 749–761. https://doi.org/10.1037/0021-9010.91.4.749

Motowidlo, S. J., Lievens, F., & Ghosh, K. (2018). Prosocial implicit trait policies underlie performance on different situational judgment tests with interpersonal content. *Human Performance*, *31*(4), 238–254. https://doi.org/10.1080/08959285.2018.1523909

Mussel, P., Gatzka, T., & Hewig, J. (2018). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment*, *34*(5), 328–335. https://doi.org/10.1027/1015-5759/a000346

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin*, *31*(2), 166–180. https://doi.org/10.1177/0146167204271418

Olaru, G., Burrus, J., Maccann, C., Zaromb, M. F., Wilhelm, O., & Roberts, D. R. (2019). Situational judgment tests as a method for measuring personality: Development and validity evidence for a test of dependability. *PLoS One*, *14*(2), e0211884. https://doi.org/10.1371/journal.pone.0211884

Olaru, G., Jankowsky, K., Mussel, P., & Mazziotta, A. (2019, September). *Situational judgment measures of personality. Response formats and scoring procedures*. Paper presented at the 15th DPPD conference, Dresden, Germany. https://osf.io/g3cqx/

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*(4), 335–353. https://doi.org/10.1080/08959285.2012.703732

Oostrom, J. K., de Vries, R. E., & de Wit, M. (2018). Development and validation of a HEXACO

situational judgment test. *Human Performance*, *32*(1), 1–29. https://doi.org/10.1080/08959285.2018.1539856

Parrigon, S., Woo, S. E., Tay, L., & Wang, T. (2016). CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of Personality and Social Psychology*, *112*(4), 642–681. https://doi.org/10.1037/pspp0000111

Ployhart, R. E. (2006). The predictor response process model. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 83–105). Lawrence Erlbaum Associates Publishers.

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), 1–16. https://doi.org/10.1111/1468-2389.00222

Rauthmann, J. F. (2012). You say the party is dull, I say it is lively: A componential approach to how situations are perceived to disentangle perceiver, situation, and perceiver× situation variance. *Social Psychological and Personality Science*, *3*(5), 519–528. https://doi.org/10.1177/1948550611427609

Rauthmann, J. F. (2015). Structuring situational information. A road map of the multiple pathways to different situational taxonomies. *European Psychologist*, *20*(3), 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., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 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*(2), 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*(2), 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*(3), 363–381. https://doi.org/10.1002/per.1994

Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist*, *51*(3–4), 305–316. https://doi.org/10.1080/00461520.2016.1208094

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*(2), 464–480. https://doi.org/10.1037/a0038098

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*(3), 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. (2020). 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*, *93*(2), 472–494. 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?* 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

Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2013). BEFKI GC-K: Eine Kurzskala zur Messung kristalliner Intelligenz. *Methoden, Daten, Analysen (mda)*, 7, 153–181. https://doi.org/10.12758/mda.2013.010

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*(3), 445–494. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Steyer, R., Ferring, D., & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*(2), 79–98.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

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*(4), 397–423. https://doi.org/10.1006/jrpe.2000.2292

Torres, W. J., & Beier, M. E. (2016). It's time to examine the nomological net of job knowledge. *Industrial and Organizational Psychology*, *9*, 51–55. https://doi.org/10.1017/iop.2015.116

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, *23*(6), 466–470. https://doi.org/10.1177/0963721414550706

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 295–322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

Appendix A

Developing a Short-Form Situational Judgment Test to Assess Implicit Trait Policies for Agreeableness

This article was first published as preprint on the Open Science Framework preprint server:

Freudenstein, J.-P., & Krumm, S. (2020). *Developing a short-form situational judgment test to assess implicit trait policies for agreeableness*. https://doi.org/10.31219/osf.io/kax7n

Appendix A

Developing a Short-Form Situational Judgment Test to Assess Implicit Trait Policies for Agreeableness

Jan-Philipp Freudenstein & Stefan Krumm
Freie Universität Berlin

Implicit Trait Policies (ITPs) are defined as implicit beliefs about the effectiveness of behaviors that express a certain trait. They are typically assessed with Situational Judgment Tests (SJTs). However, such tests often lack sufficient psychometric properties. In this study (N = 133), we aimed at developing a short-form of an SJT to assess ITPs for agreeableness. Results showed, that the six-item short-version had superior model fit when compared to the original test while maintaining the same correlation to self-reported personality. Overall, the short-form is suitable for future application. Limitations and future research perspectives are discussed.

Implicit Trait Policies (ITPs) are defined as implicit beliefs about the effectiveness of behaviors that express a certain trait (Motowidlo et al., 2006; Motowidlo & Beier, 2010). For instance, the general belief that agreeable behavior is always more effective than disagreeable behavior, independent of the situational context, constitutes an ITP for agreeableness (Lievens, 2017; Lievens & Motowidlo, 2016). Initially, the concept of ITPs was used to explain the relation of personality and Situational Judgment Test (SJT) scores as persons with a higher trait expression tend to have higher ITPs for that same trait (Crook et al., 2011; Kell et al., 2010; Motowidlo et al., 2006) and will thus chose trait-related behavior more often as a response in SJT items. In fact, several studies confirmed an empirical link between ITPs and personality (e.g., Martin-Raugh et al., 2016; Motowidlo et al., 2006).

ITPs are closely related to the method of SJTs (Lievens, 2017; Lievens & Motowidlo, 2016; Motowidlo et al., 2006, 2009, 2016). Lievens and Motowidlo (2016) argued, that test-takers utilize ITPs when responding to Situational Judgment Test items. Specifically, when individuals lack knowledge for specific situations, they fall back to general beliefs about the effectiveness of behaviors (i.e., ITPs). Indeed, Motowidlo and Beier (2010) showed that a scoring key developed by students without prior job experiences predicted relevant job criteria. Within traditional SJTs, ITPs can be operationalized as the relation of a person's effectiveness rating and the degree of trait expression of a given response option (Lievens, 2017). However, some SJTs have been developed in a way that the "true" effectiveness and the trait expression of response options align perfectly (i.e., effective response options reflect high trait expression and vice

Correspondence concerning this manuscript should be addressed to Jan-Philipp Freudenstein, jpfreudenstein@gmail.com. We are thankful to Stephan J. Motowidlo for providing us with the original test version. All data and code are available on the Open Science Framework (osf.io/kax7n)

versa; Motowidlo et al., 2009, 2016). However, SJTs typically come with insufficient psychometric properties (Guenole et al., 2017). For instance, the construct validity of SJTs has recently been described as "hot mess" (McDaniel et al., 2016, p. 47). This methodological shortcoming is even more severe, when the assessment of a novel construct is proposed (i.e., ITPs). Thus, this study aims to develop a short and psychometrically sound measurement of ITPs for Agreeableness. We apply Ant Colony Optimization (ACO), a heuristic method, which has been demonstrated to be superior in the construction of short scales when compared to several other approaches (Olaru et al., 2015). The new short scale of ITPs for Agreeableness may be fruitful for scrutinizing underlying processes of traditional SJTs.

Methods

Sample

An a-priori power analysis ($\alpha = .05$; $1-\beta = .80$) with SemPower (Moshagen & Erdfelder, 2016) revealed a required sample size of $n = 109$ to detect an RMSEA = .05. A total of $N = 133$ (72 female) individuals took part in the study. All data was collected online via the panel provider prolific.co. As reimbursement, participants received £2 for an average time to complete of $M = 18.76$ ($SD = 9.88$) minutes. On average, participants were $M = 39.67$ ($SD = 11.64$; range: 21-67) years old. All participants worked at least part-time in a job that required frequent contact to co-workers or customers with $M = 36.21$ ($SD = 9.31$) weekly work hours and average work experience of $M = 19.68$ ($SD = 11.99$) years. We checked for careless responding (Meade & Craig, 2012) by asking participants whether we should use their data for analyses, checking for zero within-person variance in responses, and significant Mahalanobis distances ($p < .001$). We excluded no participants for data analyses.

Measures

Implicit Trait Policies

We applied an SJT that has been developed to assess ITPs for Agreeableness (Motowidlo et al., 2006). This SJT consists of 22 job-related situation descriptions with four response options. Each response option reflects either agreeable or disagreeable behavior. We asked participants to rate the effectiveness of each response option in the given situation on a 7-point scale (1 = very ineffective; 7 = very effective). We scored the SJT by recoding all response options that reflect disagreeable behavior and computing mean scores for each situation descriptions. Reliability of this test was $\alpha = .89$ and $\omega = .90$.

Agreeableness

We also asked participants for self-report ratings on three pairs of adjective markers assessing agreeableness. We used those three pairs for which Wood, Nye and Saucier (2010) reported the highest average convergent correlations to three different trait assessments of agreeableness (kind-hearted, caring; giving, generous; rude, inconsiderate). Participants rated each adjective pair on a 7-point rating scale with regard to how they see themselves at work (1 = very uncharacteristic or untypical of me; 7 = very characteristic or typical of me). Reliability of this scale was $\alpha = .73$ and $\omega = .73$.

Data Analyses

We used ACO to develop a valid short form of the SJT (Leite et al., 2008; Olaru et al., 2015). ACO is a heuristic method that optimizes solutions based on defined criteria instead of choosing the best out of

all possible solutions. ACO is implemented in the mmas function of the R-package stuart (version 0.7.3; Schultze, 2019). We constructed a six-item version of the SJT by optimizing towards latent model fit (i.e., RMSEA and SRMR) and composite reliability of the short version. In addition, the internal consistency of response options within situation descriptions was used as heuristic information for the selection of each item. Heuristic information influences the selection probability of items independently from the quality of a certain solution (Schultze, 2017). That is, selection probability of an SJT scenario was higher with increasing internal consistency of response options within the same scenario. We determined the final solution based on 10 runs of the ACO algorithm. All confirmatory factor analyses (CFA) were estimated with lavaan (version 0.6-3; Rosseel, 2012). Due to missing values for two participants, we applied full information maximum likelihood estimation (Enders & Bandalos, 2001). All data and R code are available on the Open Science Framework (https://osf.io/qc8x2/).

Results

The CFA for the complete SJT with 22 items did not yield an acceptable fit; $\chi^2(209) = 392.58$, $p < .001$; CFI = .80; RMSEA = .08; SRMR = .07. Despite the good overall reliability ($\omega = .90$), internal consistencies for the four response options to every situation description were quite low on average ($M = .38$; $SD = .25$). Bivariate correlations are depicted in Table 1.

All 10 ACO runs converged to the same six-item short version (see Appendix A). This model had an excellent fit; $\chi^2(9) = 7.31$, $p = .605$; CFI = 1.00; RMSEA = .00; SRMR = .03. The internal consistency of this short form was good but somewhat lower in comparison to the original test ($\omega = .81$). Internal consistencies of response options within single SJT items were higher when compared to the original version ($M = .49$; $SD = .17$). A comparison of bivariate correlations to self-report Agreeableness did not yield a significant difference between the original and short version ($z = 1.02$, $p = .309$).

Table 1

Descriptive Statistics and Bivariate Correlations

		M (*SD*)	1.	2.	3.
1.	**ITP-SJT**	4.83 (0.49)	-		
2.	**ITP-SJT short**	4.93 (0.67)	.92	-	
3.	**Agreeableness**	5.69 (1.08)	.26	.23	-

Notes. $n = 131$. ITP = Implicit Trait Policy. SJT = Situational Judgment Test.

Discussion

The aim of this study was to develop a short measure of ITPs for Agreeableness with good psychometric properties. Results showed that the newly developed short version has superior model fit in comparison to the complete SJT. Moreover, the short version retained similar properties with regard to the relation to self-report Agreeableness and was highly correlated with the original version. Thus, the short-form is suitable for future application as it complies with standards required for psychological assessment. Overall, results are in line with research showing that ACO can be a superior tool for the construction of valid psychological assessments (Olaru et al., 2015).

Thus far, advances in research on ITPs have been primarily made on a conceptual and theoretical level (Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010). Although ITPs have been proposed as underlying construct of SJTs, little evidence about the relation of ITPs and SJT performance exists (cf., Motowidlo et al., 2018; Oostrom et al., 2012). Therefore, future research is needed that further dives into scrutinizing the relevance of ITPs for SJT performance. This research provides a psychometrically sound measure of ITPs for Agreeableness that may contribute to this undertaking. It was especially required when considering the fact that the assessment of ITPs relies on the same methodological approach as traditional SJTs, which typically has psychometric shortcomings (Guenole et al., 2017; McDaniel et al., 2016).

Nevertheless, this short ITP measure should be applied with care. First, the optimized short scale was not cross-validated on a second sample. Sampling error and context effects of items that were not selected for the short scale could affect the suitability of the derived solution. Future research should pay close attention to the model fit of the short scale. Second, construct validity of the SJT was not assessed. Although we assessed self-reported Agreeableness, we did not measure ITPs for Agreeableness with different tests. Thus, the derived short scale can only be understood as psychometrically superior version of the original test. Further research is needed to assess the construct validity of this short version and ITPs in general, especially since research on ITPs lacks such studies (cf., Motowidlo et al., 2018).

References

Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance using single-response situational judgment tests. *International Journal of Selection and Assessment*, *19*(4), 363–373. https://doi.org/10.1111/j.1468-2389.2011.00565.x

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. https://doi.org/10.1207/S15328007sem0803_5

Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing*, *17*(3), 234–252. https://doi.org/10.1080/15305058.2017.1297817

Kell, H. J., Rittmayer, A. D., Crook, A. E., & Motowidlo, S. J. (2010). Situational content moderates the association between the big five personality traits and behavioral effectiveness. *Human Performance*, *23*(3), 213–228. https://doi.org/10.1080/08959285.2010.488458

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*(3), 411–431. https://doi.org/10.1080/00273170802285743

Lievens, F. (2017). Assessing personality-situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*(5), 424–440. https://doi.org/10.1002/per.2111

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*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Martin-Raugh, M. P., Kell, H. J., & Motowidlo, S. J. (2016). Prosocial knowledge mediates effects of agreeableness and emotional intelligence on prosocial behavior. *Personality and Individual Differences*, *90*, 41–49. https://doi.org/10.1016/j.paid.2015.10.024

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*(1), 47–51. https://doi.org/10.1017/iop.2015.115

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 54–60. https://doi.org/10.1080/10705511.2014.950896

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*(2), 321–333. https://doi.org/10.1037/a0017975

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology*, *24*(3), 281–288. https://doi.org/10.1007/s10869-009-9106-4

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance*, *29*(4), 331–346. https://doi.org/10.1080/08959285.2016.1165227

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology*, *91*(4), 749–761. https://doi.org/10.1037/0021-9010.91.4.749

Motowidlo, S. J., Lievens, F., & Ghosh, K. (2018). Prosocial implicit trait policies underlie performance on different situational judgment tests with interpersonal content. *Human Performance*, *31*(4), 238–254. https://doi.org/10.1080/08959285.2018.1523909

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*(4), 335–353. https://doi.org/10.1080/08959285.2012.703732

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schultze, M. (2017). *Constructing subtests using ant colony optimization* [Doctoral dissertation, Freie Universität Berlin]. http://doi.org/10.17169/refubium-622

Schultze, M. (2019). *stuart: Subtests using algorithmic rummaging techniques*. https://cran.r-project.org/package=stuart

Wood, D., Nye, C. D., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive trait markers from the english lexicon. *Journal of Research in Personality*, *44*(2), 258–272. https://doi.org/10.1016/j.jrp.2010.02.003

Appendix

Situational Judgment Test Items

Instruction

This questionnaire asks for your opinions about how to deal with people at work. It contains descriptions of some awkward or difficult interpersonal situations that might happen in large work organizations and alternative actions that a person could take to deal with them. It asks you how effective the actions are for dealing with each situation.

Each interpersonal situation described in this questionnaire has four action alternatives listed directly below it. Please write a number from 1 to 7 on the lines next to the action alternatives to show how effective you think they are, where...

7 = very effective	4 = neither effective	3 = slightly ineffective
6 = somewhat effective	nor ineffective	2 = somewhat ineffective
5 = slightly effective		1 = very ineffective

Item	Situation Description	Response Options
1	Your recently appointed superior has had many disagreements with you and your colleagues, and usually supports her position by indicating that her view is "how things were done" in her previous job environment. In a meeting with you and your team, after you propose a way to solve a procedural problem, she said, "no, we'll do it the way I've always done it". You should…	(a) tell her that in this case her suggestions are simply wrong. *(-)* (b) meet with her privately and explain that you feel she is not accepting any input from the team and this is hurting morale. *(+)* (c) tell her that you will be looking for a new assignment if this is the way you will be working together. *(-)* (d) meet with her privately and explain the effect she is having on you and the entire team. *(+)*
4	You and your colleague, who is at the same level in the organization, "share" the same clerk. One day the clerk takes you aside and tells you that your colleague told her that his requests should take priority over yours. You should...	(a) meet with your colleague to discuss the situation and suggest setting up a schedule for the clerk so you can coordinate your tasks more efficiently. *(+)* (b) ask your colleague if he is working on a special project, deadline, or is planning to be away so that you can find a compromise that would suit you both. *(+)* (c) tell the clerk to ignore the comment made by your colleague and to treat both of your requests equally. *(-)* (d) get a written statement from your manager of what the clerk's priorities are and show it to your colleague. *(-)*

Appendix (continued)

11	You suspect that your manager has been taking credit for documents that you have prepared and ideas that you have generated. One afternoon you notice him attaching his business card to a presentation that you prepared. You should…	(a) speak with your manager and tell him that the lack of recognition makes you feel unmotivated at work. *(+)* (b) tell your manger that you think his behavior is unethical and that you will be filing a complaint. *(-)* (c) tell your manager how much work you put into the presentation and that you would appreciate the recognition for it. *(+)* (d) tell your manager that if he doesn't stop attaching his business cards to your presentations, you will have no alternative but to report his actions. *(-)*
13	You are the newest member of a project team and you are at your first team meeting. You have just started presenting what you think is a good idea when you are interrupted by another member who tells you that, because you are so new and inexperienced, you should sit back quietly and learn. You should…	(a) tell the individual that the group, not one person, should judge the merit of your idea. *(-)* (b) speak privately with the individual after the meeting about how their comments made you feel. *(+)* (c) acknowledge that you are new and ask the entire group to listen to you as an equal member of the team. *(+)* (d) tell the individual privately that you expect an apology in front of the other team members at the next meeting. *(-)*
20	You are in charge of a meeting with six people from other departments. One of them has a very blunt way of announcing that something that was said is stupid or that somebody's idea just won't work. By the time the meeting is half over, he has done this twice in connection with remarks made by two different participants. The meeting is scheduled to continue for another thirty minutes. You should…	(a) during a break or after the meeting, explain to him that you appreciate his point of view, but that his comments are hurting the other coworkers. *(+)* (b) during the meeting, tell him to keep his rude comments to himself or he won't have a job anymore. *(-)* (c) during a break or after the meeting, tell him that his comments were hurting group participation, and ask him to phrase his criticisms differently. *(+)* (d) during the meeting, ask him to leave the meeting. *(-)*

Appendix (continued)

21	You have recently been promoted to a management job. On the second day in your new position, one of your new subordinate team members comes into your office and tells you she thinks she should have been promoted to your position instead of you because she has more seniority and technical experience. You should...	(a) listen to her concerns, letting her know that her experience is valued and where possible you would like to help with her career development. *(+)* (b) acknowledge her technical experience and affirm her value to the team. *(+)* (c) explain to her that you are not responsible for the selection process and that she should be taking her concerns to the appropriate people elsewhere. *(-)* (d) tell her that you went through the proper channels to get the promotion and that despite her frustration she should treat you as her supervisor. *(-)*

Notes. Scoring instructions in parentheses. + Response rating is included in the sum score. – Reverse-coded response rating is included in the sum score. These items can be applied for research purposes by citing this manuscript and Motowidlo et al. (2006). Stephan J. Motowidlo holds the copyright on all items.

Appendix B

English Translation of the Teamwork Situational Judgment Test (SJT-TW)

This article has been submitted for publication:

Freudenstein, J.-P., Remmert, N., Reznik, N., & Krumm, S. (2020). *English translation of the teamwork situational judgment test (SJT-TW)* [Manuscript submitted for publication to Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)].

1 Overview

Abstract

The Teamwork Situational Judgment Test (SJT-TW; Gatzka & Volmer, 2017) consists of 12 situation descriptions with four response options and assesses individual teamwork effectiveness. So far, only a German version of this test exists. To translate the SJT-TW to English, we utilized the TRAPD procedure (Harkness, 2003). TRAPD is an acronym for several steps needed to produce high-quality translations of questionnaires, namely translation, review, adjudication, pretesting and documentation. Results from a pilot study provide preliminary evidence for item and test score properties of the translated test when compared to the original German version.

Keywords

Title: English Translation of the SJT-TW

Author: Freudenstein, Remmert, Reznik & Krumm

In ZIS since: 2020

Number of Items: 12

Reliability: Cronbach's alpha = .52

Validity: no validity evidence for the translated version

Construct: teamwork effectiveness

Catchwords: teamwork, team spirit, SJT

Language Documentation: English

Language Items: English

URL Data Archive: https://osf.io/zgank/

Item(s) used in Representative Survey: no

Status of Development: tried

Survey Mode: paper-pencil, CASI

Duration: 10 minutes

2 Instrument

Please note that the scoring key may be requested by sending the signed Data Use Agreement to zis@gesis.org or the original test authors (Thomas Gatzka and Judith Volmer; see Scoring Key Authorization). The following instruction describes how the test was applied in this study. In contrast to the German test version, we only asked participants to pick the best out of four response options. The German version additionally asks for the worst response option. We chose this shortened instruction to reduce the test duration in an initial pilot study. The complete instruction is presented within square brackets and in grey font colour.

Instruction

Below, 12 situations are described as they may occur in the occupational daily routine of teams or working groups. For each situation, four different behavioural options are presented.

Please pick the *most* [and *least*] suitable behaviour for *each* situation.

For some situations, it may be difficult for you to decide as certain details are not specified, you did not experience a similar situation before, or you consider some options very similar. However, please choose the alternatives that you generally take for the best [and worst] solution.

Please always select the *best* [and the *worst*] option for each situation. [Please do not indicate the *same* answer as the best *and* worst solution.]
Please do not skip any situation.

Example

Your team has a task that is fundamentally different to previous tasks and covers completely new aspects. In addition, it is very likely that aspects of the task will change in the medium term.

What should your team do [and not do] in such a situation?

a) Some members of the team do not assist with the task to stay flexible. (X)
b) All aspects of the tasks are assigned to several competent members of the team. ( )
c) The team asks a supervisor to assign task aspects. ( )
d) Task aspects are assigned as needed in regular meetings. ( )

Items

Table 1

Items of the English Version of the SJT-TW

Situation	Items	Given Answers
1	You are temporarily subjected to personal stress that also affects your occupational activity. A briefly acquainted colleague asks you about the reason for your decline in performance and offers help with your task.	What should you do and not do in such a situation? (a) You confirm an increase in personal stress and accept the help. (b) You do not mention your personal problems, but accept the help. (c) You explain your specific situation and ask for help with your task. (d) You thank your colleague for the feedback but politely decline the offer.
2	Your team has clearly allocated all areas of responsibility. However, you incidentally notice that some team members in another area are challenged by a task that you have experience in.	What should you do and not do in such a situation? (a) You tell your colleagues about your experience and offer advisory support. (b) You mention your expertise and offer active support. (c) You ask whether your experience or advice is desired and if so, when. (d) Your respect your colleagues' responsibilities and stay out of it.
3	Your team has made a lot of progress working on a complex task when some unforeseen developments occur. Therefore, your tediously achieved results are no longer completely up to date.	What should your team do and not do in such a situation? (a) Slight shortcomings will be tolerated due to the advanced progress of the work. (b) The team asks the customer or superior for their assessment of the situation. (c) Team members immediately discuss possible consequences in a meeting. (d) Changes are retrospectively implemented through intensive additional work.
4	You notice a sudden but continuous decrease in performance in one of your team members, whom you have experienced as competent and reliable. Other sources report that this colleague currently has some personal problems.	What should you do and not do in such a situation? (a) You and the other team members discuss how to support this colleague. (b) You respect your colleague's privacy and do not get involved in private matters. (c) You help your colleague without asking questions. (d) You ask your colleague if they want to talk about their problems.

Table 1 (continued)

5	Some of your colleagues discuss various aspects of a team task during a meeting. Your area of responsibility is not the focus of their discussion, which is why you hold back and refrain from partaking in it.	What should you do and not do in such a situation? (a) You mentally prepare the discussion points that you wanted to address. (b) You use the opportunity to broaden your knowledge about other parts of the task. (c) You carefully steer the conversation towards a more familiar topic that you can engage in. (d) You attentively look for information that could be important for your area of responsibility.
6	You are transferred to an already existing team. During a brief introduction, your contact person tells you that all team members have their own area of responsibility. Without providing any further details, your contact person instructs you on your own area of responsibility.	What should you do and not do in such a situation? (a) You limit your questions to your area of responsibility, as nothing else should concern you. (b) You hold back your curiosity and carefully listen to your contact person's explanations. (c) You decide to become familiar with the other areas of responsibility on your own after the conversation. (d) You ask for the basic workflows and interdependencies in the team.
7	You have to inform another team member about a complex issue from your area of responsibility. It is of utmost importance for your team's success that the other person takes note of your concern and that no uncertainties are left.	What should you do and not do in such a situation? (a) You prepare a timesaving summary that you personally deliver. (b) You send a detailed report and ask for an acknowledgement of receipt. (c) You arrange a personal meeting with the other team member. (d) You send a message and explicitly request the other team member to contact you if any uncertainties are left.
8	You incidentally notice that another team member struggles to finish their work on time. You have already completed your tasks. However, you want to double check your work and, if necessary, improve some details before the deadline.	What should you do and not do in such a situation? (a) You ask the team member in a confidential conversation whether they need help. (b) You contain yourself because you do not want the team member to appear incompetent. (c) You carefully address your observation in the next team meeting. (d) You finish your own tasks first before offering your help.

Table 1 (continued)

9	Together with your team members, you are setting objectives for each member for an upcoming task.	What should your team do and not do in such a situation? (a) The team sets objectives that are positive, clearly defined and easily verifiable. (b) The team sets objectives that are specific, challenging and agreed upon by the whole team. (c) The team sets objectives that are moderately difficult and comprehensible to the whole team. (d) The team sets objectives that are easily attainable, open and flexible concerning time management.
10	You are working on a task that is mainly in your area of responsibility. When you present your intended procedure during a meeting, some team members from other areas speak up and add suggestions for changes and adaptations.	What should you do and not do in such a situation? (a) You take note of suggestions and discuss them with everyone involved. (b) You reflect on what changes might be sensible and ask for details. (c) You politely point out that you have a better overview of the task due to your expertise. (d) You try to include as many of the suggested changes as possible in your plan.
11	Due to external circumstances, your team was unable to finish an important task on time. Since every team member has given their very best, there is considerable disappointment. When a new task comes up, you notice that low morale and poor motivation are impairing the team.	What should you do and not do in such a situation? (a) You remind your team of past successes to spark new motivation. (b) You address your concerns in front of the whole team and encourage a discussion. (c) You ask for a team meeting to put the failure behind you. (d) You give the other team members the time to regain their motivation.
12	Together with your team members, you are planning how to tackle an upcoming task. The team's success in mastering this challenge depends on several factors, some of which are difficult to predict.	What should your team do and not do in such a situation? (a) The team discusses all possible developments in advance and works out a strategy for each of them. (b) The planning proceeds in small steps in order to allow quick adaptations. (c) The team waits with further planning until all uncertainties are eliminated. (d) The team focuses especially on currently available facts for the planning.

Response specifications

The answers are given in a forced-choice format. The best and the worst solution has to be identified. The respondent's task is to indicate how they (or the entire team) should behave.

A computer-assisted format usually forces test-takers to give two responses. However, in a paper-pencil format, this may not be obvious to test-takers. Thus, the following sentence should be included in the test instruction subsequent to the sentence "Please always select the best and the worst option for each situation":

"Please select exactly two response options for each situation. Mark (+) for the best solution and (-) for the worst solution."

Scoring

For each scenario there is a predefined best and worst solution, which can be taken from the scoring key. The scoring key may be requested by sending the signed Data Use Agreement to zis@gesis.org or the original test authors (Thomas Gatzka and Judith Volmer; see Scoring Key Authorization). If test-takers correctly choose the best solution, the response is coded as "1". If test-takers correctly choose the worst solution, the response is also coded as "1". If test-takers select the best solution as their worst solution or vice versa, the responses are scored as "-1". All remaining responses are scored as "0". Values are added up for each scenario. Thus, each scenario can have values from -2 to +2. To obtain a score for the total test, values across scenarios are added up to an unweighted sum score. Test scores may also be obtained separately for best and worst responses across scenarios.

Adequate methods may be applied to deal with missing values (i.e., multiple imputation; full information maximum likelihood).

Application field

This test should be applied to assess knowledge about teamwork effectiveness in research settings (given the lack of validity evidence for the translated version of the test, we do not encourage its use beyond research settings). This test can be applied independently from actual teams or team tasks. For instance, the original development study (Gatzka & Volmer, 2017) validated this test with a student sample as well as a sample of employees. It may be particularly useful for teamwork research (see Gatzka & Volmer, 2017). The test may be applied in a computer-assisted or a paper-pencil self-administered questionnaire format. On average, participants took 6.57 minutes (*SD* = 1.63) to complete the test with the shortened instruction. Hence, participation will take

approximately 10 minutes if test-takers are asked to pick both the best and worst response options.

3 Theory

The reported test is a translation of the German Situational Judgment Test for Teamwork (SJT-TW; Gatzka & Volmer, 2017). SJTs are popular tools in personnel selection and are traditionally defined as low-fidelity simulations (Motowidlo et al., 1990). Most SJTs consist of written situation descriptions and several behavioural response options of which test-takers chose the most similar to how they should or would behave in the given situation (McDaniel & Nguyen, 2001). As such, they sample knowledge about effective behaviours in relevant situations for work-related criteria (Motowidlo et al., 1990; Weekley et al., 2015). Meta-analyses confirmed the predictive power of SJTs for job performance criteria (Christian et al., 2010; McDaniel et al., 2001, 2007).

Effective teamwork can be best described as a set of various behaviours rather than a single, narrow construct (Salas et al., 2005; Rousseau et al., 2006). Accordingly, Gatzka and Volmer (2017) identified SJTs as suitable tool for the assessment of teamwork effectiveness. These authors demonstrated that the SJT-TW correlated with measures of teamwork skills and even predicted supervisor-rated contextual and teamwork performance. Overall, the original version of the test was well in line with contemporary conceptualizations of teamwork effectiveness and thus a valuable tool for teamwork research and personnel selection (Gatzka & Vollmer, 2017).

Beyond the intended use of the SJT-TW for teamwork research, the test may be useful for research on the underlying psychological processes of SJTs. Despite the well-established criterion-related validity of SJTs, it remains unclear why SJTs work as assessment methods (e.g., Freudenstein et al., 2020; Lievens & Motowidlo, 2016; McDaniel et al., 2016; Schäpers et al., 2019). For instance, the role of situations for test-takers' responses to SJT items is subject to debate. Some argue in favour of processes that are similar to those underlying behaviour in real-life situation, while others advocate context-independent constructs (e.g., Freudenstein et al., 2020; Lievens & Motowidlo, 2016; Schäpers et al., 2019). However, the number of SJTs that are available to research is limited. Thus, an English translation of the SJT-TW would further enable research about underlying processes of SJTs.

4 Scale development

Item generation and translation

Gatzka and Volmer (2017) integrated results from two reviews on teamwork to develop a working model and to identify core elements of team effectiveness (Salas et al., 2005; Rousseau et al., 2006). Furthermore, they considered two models that have already been implemented in test procedures (O'Neil, Jr., et al., 1997; Stevens & Campion, 1994) as well as further reviews on team processes and team efficacy (Kozlowski & Ilgen, 2006; Marks et al., 2001; Mathieu et al., 2008). The working model consisted of 30 behaviours particularly relevant for teamwork success.

Gatzka and Vollmer (2017) used these behaviours to develop a Situational Judgment Test. Their final test consists of 12 hypothetical situations or scenarios that reflect a problem concerning teamwork and four behavioural response options for each situation. Test-takers are asked to indicate the best and worst solution for each situation. The SJT showed substantial correlations with related constructs and job-related criteria.

To translate the SJT-TW to English, we utilized the TRAPD procedure (Harkness, 2003). TRAPD is an acronym for several steps needed to produce high-quality translations of questionnaires, namely translation, review, adjudication, pretesting and documentation. We created two independent translations of the SJT-TW. The overall aim was to retain as much original item content and structure as possible. Both translators were fluent in spoken and written English and had expertise in SJT research. However, both translators were neither native speakers nor professional translators. The first author reviewed both translations and merged them into a single version. Afterwards the translators revised this version with regard to word flow and completeness of the original item content. Two independent native speakers then additionally reviewed this revised test version. All changes were adopted accordingly. Next, a senior researcher with high expertise in psychological assessment and SJT research made final changes to the translation.

We additionally pilot-tested the translated SJT with a small sample to gauge whether test-takers understood all items and to inspect preliminary response patterns. We instructed participants to pick the response options that best resembles what they should do in each of the 12 scenarios. To reduce the duration to participate, we did not instruct participants to pick the response option that resembles the worst solution. This is contrary to the original test format. We scored responses with “1” if they reflected the most effective response, with “-1” if they reflected the most ineffective response, and all remaining responses with “0”. Please note that interpretations of these results are only preliminary and should be made with caution due to the small sample size. Data was analysed with R (version 3.6.1; R Core Team, 2019) and the R package psych (version 1.8.12; Revelle, 2018).

Sample

Data for the English version of the SJT-TW was collected in 2019 from the following

convenience sample from the United States: *N* = 20 native speakers (American English) from Amazon MTurk; sex: = 40% female; age: *M* [min; max] = 35.25 [25; 53], *SD* = 9.21. Most participants (75%) were gainfully employed during the time of data collection. Participants had either an undergraduate (45%) or graduate degree (20%) or received vocational training (5%). The remaining 30% of the sample graduated high school. Test-takers received $1 for participation. No a-priori power analysis was conducted, as this was a pilot study. No missing values occurred.

Item analyses

Table 2 presents item parameters for the 12 SJT items. Item distributions were somewhat similar to those of the German version (Gatzka & Volmer, 2017). The range of item total correlations was also comparable between the German and the English version of the SJT-TW, with a slightly higher mean of item-total correlations for the English version (r_{it} = .22 vs. .17). The internal consistency of SJTs is typically low (Catano et al., 2012; Kasten & Freund, 2016). Thus, small item-total correlations were to be expected. However, item 11 showed a negative item-total correlation. This may be due to the small sample size of this pilot study. Nevertheless, if a negative item-total correlation persists in future applications, this item should be excluded from further analyses.

The reported item-total correlations presume a single factor structure of the SJTs. This is in line with recommendations by Gatzka and Volmer (2017). However, these authors also proposed a two-factor structure of the SJT-TW (Factor 1: Items 2, 3, 5, 6, 7, 9, 10, 12; Factor 2: Items 1, 4, 8, 11). Gatzka and Volmer (2017) argued that this factor structure can only be interpreted as preliminary evidence due to the small number of items and low internal consistencies of the two factors. They concluded that only a total test score should be calculated. An investigation of the factor structure of the translated SJT-TW was not sensible due to the small sample size of *N* = 20.

Table 2

Means, Standard Deviations, Skew, Kurtosis and Item-Total-Correlations of the Manifest Items

	M	*SD*	Skew	Kurtosis	r_{it}
Item 1	0.70	0.47	-0.81	-1.41	0.43
Item 2	0.00	0.65	0.00	-0.74	0.39
Item 3	0.25	0.55	0.11	-0.60	0.04
Item 4	-0.20	0.41	-1.39	-0.07	0.13
Item 5	0.40	0.60	-0.34	-0.95	0.49
Item 6	0.40	0.50	0.38	-1.95	0.25
Item 7	0.25	0.91	-0.47	-1.68	-0.02
Item 8	0.20	0.70	-0.25	-1.06	0.38
Item 9	0.25	0.72	-0.36	-1.12	0.11
Item10	0.15	0.67	-0.15	-0.93	0.32
Item 11	0.25	0.44	1.07	-0.89	-0.22
Item 12	0.15	0.75	-0.22	-1.27	0.29

Note. Scale ranging from -1 to 1 as test-takers only were asked to pick the best response option, $N = 20$.

5 Quality criteria

Objectivity

The English translation of the SJT-TW is a standardised psychological instrument like the German original SJT-TW. Each test-taker receives the same instruction, items and response options. The answers are evaluated by means of a solution key. Hence, objectivity of application and evaluation is assured. Due to the ambiguous factor structure of the SJT-TW, test scores should be interpreted with care. Rather than allocating psychological meaning to tests scores, sum scores of the SJT-TW should be interpreted as indicators that correlate with various constructs such as job performance and team skills. This is not unique to this specific test but rather representative for most SJTs (see McDaniel et al., 2016).

Reliability

The reliability of the scale was determined by internal consistency estimator Cronbach's alpha. Coefficient alpha of the 12 SJT items was α = .52. Although this internal consistency is insufficient, it is similar to the sample of employees in the original validation study (α =.44; Gatzka & Volmer, 2017). Moreover, the internal consistency of the SJT-TW is in line

with meta-analyses on the internal consistency of SJTs (Catano et al., 2012; Kasten & Freund, 2016). These low values generally reflect the ambiguous factor structure of SJTs.

Validity

Based on results from this very small sample, we tentatively conclude that the overall scale worked very similar to the German version and did not cause any major inconsistencies. Still, a proper validation study is needed before using the English SJT-TW beyond research settings. We consider the current version as a research version, which should not be used in high stakes settings.

Descriptive statistics (scaling)

The test sum score had a mean of 2.80 (*SD* = 3.02) with a skewness of -1.04 and a kurtosis of 0.37. Thus, participants chose on average more correct than incorrect response options This result is in line with the German test version (Gatzka & Volmer, 2017).

Further quality criteria

The test processing takes about 10 minutes, which indicates that the test is a very economical instrument. Research also suggests that SJTs are less susceptible to faking behaviour, especially when compared to personality self-reports (Kasten et al., 2018).

6 Literature and data sources

Further literature

Gatzka, T., & Volmer, J. (2017). Situational Judgment Test für Teamarbeit (SJT-TA). In *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis249

Contact details

- Jan-Philipp Freudenstein, Freie Universität Berlin, jpfreudenstein@gmail.com

References

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*(3), 333–346. https://doi.org/10.1111/j.1468-2389.2012.00604.x

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*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Freudenstein, J.-P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. Advance Online Publication. https://doi.org/10.1111/peps.12385

Gatzka, T., & Volmer, J. (2017). Situational Judgment Test für Teamarbeit (SJT-TA). In *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis249

Harkness, J. (2003). Questionnaire Translation. In J. Harkness, F. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Wiley.

Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). *European Journal of Psychological Assessment*, *32*(3), 230–240. https://doi.org/10.1027/1015-5759/a000250

Kasten, N., Freund, P. A., & Staufenbiel, T. (2020). Sweet little lies. An in-depth analysis of faking behavior on situational judgment tests compared to personality questionnaires. *European Journal of Psychological Assessment*. *36*(1), 136–148. https://doi.org/10.1027/1015-5759/a000479

Kozlowski, S. W. J. & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, *7*(3), 77-124. https://doi.org/10.1111/j.1529-1006.2006.00030.x

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*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Marks, M. A., Mathieu, J. E. & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356-376. https://doi.org/10.5465/amr.2001.4845785

Mathieu, J. E., Maynard, M. T., Rapp, T. & Gilson, L. (2008). Team effectiveness 1997-2007. A review of recent advancements and a glimpse into the future. *Journal of Management*, *34*(3), 410-476. https://doi.org/10.1177/0149206308316061

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*(1), 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: Perspectives on Science and Practice*, *9*(1), 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*(4), 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*(1-2), 103–113. https://doi.org/10.1111/1468-2389.00167

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

O'Neil, H. F., Chung, G. K. W. K. & Brown, R. S. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil, Jr. (Eds.), *Workforce readiness. Competencies and assessment* (pp. 411-452). Erlbaum.

Revelle, W. (2018). *Psych: Procedures for Personality and Psychological Research* [Computer software]. Northwestern University. https://cran.r-project.org/pacakage=psych

Rousseau, V., Aubé, C. & Savoie, A. (2006). Teamwork behaviors. A review and an integration of frameworks. *Small Group Research*, *37*(5), 540-570. https://doi.org/10.1177/1046496406293125

Salas, E., Sims, D. E. & Burke, C. S. (2005). Is there a "Big Five" in Teamwork*? Small Group Research*, *36*(5), 555-599. https://doi.org/10.1177/1046496405277134

Stevens, M. J. & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork. Implications for human resource management. *Journal of Management*, *20*(2), 503-530. https://doi.org/10.1177/014920639402000210

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*. Advance online publication. https://doi.org/10.1037/apl0000457

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 295–322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

Appendix C

Effects of Situation Descriptions on the Construct-Related Validity of Construct-Driven Situational Judgment Tests

Schäpers*, P., Freudenstein*, J.-P., Mussel, P., Lievens, F., & Krumm, S. (2020). Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests. *Journal of Research in Personality*. https://doi.org/10.1016/j.jrp.2020.103963

*Both authors contributed equally to this work

Appendix D

Supplementary Information to Chapter 3

This supplementary information is also available online: doi.org/10.1111/peps.12385

Table S1

Data Analyses Study 1

Data Cleaning	We excluded careless responders (see Meade & Craig, 2012) based on zero within-person variance in responses (n = 1) and the 95% quantile of Mahalanobis distances across all variables (n = 12). Mahalanobis distances describe the multivariate distance between an individual's responses to all variables and the sample mean of the multivariate distribution. Thus, they indicate the extent to which a person deviates from the multivariate normal distribution. As such, Mahalanobis distances are suitable for the detecting careless responders (Meade & Craig, 2012). We additionally asked participants whether we should use their data for analyses and excluded participants who indicated that they did not understand the survey or did not pay attention (n = 8).
Data Analyses	We constructed a valid short version of the self-consciousness SJT with Ant Colony Optimization (Leite, Huang, & Marcoulides, 2008; Olaru, Witthöft, & Wilhelm, 2015) based on the original validation sample (Mussel Gatzka & Hewig, 2018). Ant Colony Optimization is a heuristic algorithm method that can be used for short scale construction. We used fit indices from confirmatory factor analysis to obtain a valid short version. Importantly, Ant Colony Optimization has been demonstrated to yield superior results compared to other methods of short scale construction (Olaru et al., 2015). However, as our research questions do not relate to scale construction and the psychometric properties of SJTs are only of secondary interest for this study, we do not provide a detailed description of the specific methodology used. For in depth information on Ant Colony Optimization for short scale construction, please see Leite et al. (2008) and Olaru et al. (2015). We checked whether the S8* and the S8-I represented the same constructs. For this purpose, we compared the manifest correlations for all SJT items' situation characteristics of both measures. This was done by restraining all correlation coefficients in a path model and testing for differences against the baseline model. For 22 out of 23 SJT items, $\Delta\chi^2$-tests revealed no significant differences in intercorrelations between the S8* and S8-I (α = .05/23) We computed pooled within-group correlations of the DIAMONDS. Since 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). Ordered logistic link models (also referred to as proportional odds models) have the advantage of adequately treating ordered dependent variables with more than two levels as categorical (Brant, 1990; Christensen, 2018). Simplified, they can be understood as set of $k - 1$ logistic regressions for k ordered categories with summary estimates based on cumulative distribution probabilities (Brant, 1990). 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 scaled perceived situation characteristics within persons and further included the grand mean centered average of each of the DIAMONDS across all SJT items as a predictor on the person level (Enders & Tofighi, 2007; see also Sherman, Rauthmann, Brown, Serfass, & Jones, 2015). This Level 2 predictor controls for general person effects that are neither due to situations nor 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). Pseudo R^2 represents improvement in model fit rather than explained variance in the dependent variable (Hemmert et al., 2016). Pseudo R^2 values tend to be smaller than R^2 values in linear regression analyses (McFadden, 1979). 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). All data and R code are available on the Open Science Framework (osf.io/6kd9h).
R Packages	*Gtheory* (version 0.1.2; Moore, 2016) *ordinal* (version 2015.6-28; Christensen, 2018) *psych* (version 1.8.12; Revelle, 2018) *userfriendlyscience* (version 0.7.2; Peters, 2018)

Table S2

Data Analyses Study 2

Data Cleaning	We excluded careless responders (see Meade & Craig, 2012) based on zero within-person variance in responses (n = 2), the 95% quantile of Mahalanobis distances across all variables (n = 40), and individual statements on whether participants recommended using their data for analysis (n = 14).
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 multi-group regression analyses for each SJT item. All participants who completed the SJT item of interest and the corresponding rating of the perceived situation characteristics were included in the analysis. In a preliminary step, we computed baseline models for which 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. Due to the categorical nature of the dependent variable, we applied the WLSMV estimator (DiStefano & Morgan, 2014). Then, 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. Similar to Study 1, R^2 for categorical data computed in *lavaan* cannot be understood as explained variance. To compute profile correlations between expert ratings of the DIAMONDS and the average situation characteristics of participants. For all mediation models, we calculated bootstrapped 95% CIs for indirect and total effects. All data and R code are available on the Open Science Framework (osf.io/6kd9h).
R Packages	*lavaan* (version 0.5-23.1097; Rosseel, 2012) *multicon* (version 1.6; Sherman, 2015) *psych* (version 1.8.12; Revelle, 2018)

Table S3

Data Analyses Study 3

Data Cleaning	We excluded careless responders (see Meade & Craig, 2012) based on the 95% quantile of Mahalanobis distances across all variables (n = 15) and individual statements on whether participants recommended using their data for analysis (n = 2). We further excluded one participant who failed to respond honestly to two bogus items (Anderson, Warner, & Spencer, 1984; Levashina, Morgeson, & Campion, 2009; Meade & Craig, 2012).
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 applied a full information maximum likelihood estimator (FIML) to appropriately deal with the missing values in the peer-rated criteria (Enders & Bandalos, 2001). Analyses for single SJT items were necessary as Hypothesis 3 specifies cross-level interactions. Thus, multilevel analysis was not appropriate due to the number of SJT items (see Hox, 2010). 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. Here, R^2 reflects explained variance due to the continuous nature of the dependent variables. We again used residual scores for the perceived situation characteristics to control for individual's general tendency to perceive multiple SJT items equally. For all mediation models, we calculated bootstrapped 95% CIs for indirect and total effects. All data and R code are available on the Open Science Framework (osf.io/6kd9h).
R Packages	*lavaan* (version 0.5-23.1097; Rosseel, 2012) *psych* (version 1.8.12; Revelle, 2018) *userfriendlyscience* (version 0.7.2; Peters, 2018)

Sample SJT Items (scoring weights in parentheses)

Personal Initiative SJT (description-independent item; Bledow & Frese, 2009)

Due to a conflict among your colleagues, the climate in your department is rather tense. You are not involved in the conflict. However, you feel disturbed in your work. The attempt of one of your colleagues to reconcile the conflict was not appreciated. What would you do?

---

least likely most likely

A. I try not to take sides and ask my colleagues to be considerate of other coworkers. **(-1)**
B. I take charge of mediating among my colleagues even if they react negatively in the beginning. **(0)**
C. I stay calm and do not let myself be bothered by the conflict. **(1)**
D. I ask my supervisor to take action. **(0)**

Self-Consciousness SJT (Mussel et al., 2016)

You are attending a public lecture together with approximately 100 other listeners. You find the talk very interesting and are keen on asking a question. What would you do?

A. I do not ask the question, because I feel inconvenient talking in front of so many people. **(1)**
B. If at all, I ask my question after the lecture, in case I meet the speaker alone. **(1)**
C. I ask my question as soon as I am sure that others will also ask questions. **(0)**
D. I ask my question as soon as the speaker briefly pauses between two sentences. **(0)**

Academic Achievement SJT (Ployhart & Ehrhart, 2003)

It's your birthday, and your friends want to take you out tonight to celebrate. Unfortunately, you have an exam tomorrow morning at 9 AM that you haven't studied for yet.

A. Go out with your friends and stay up the rest of the night studying. **(0)**
B. Go out with your friends and don't worry about studying. **(-1)**
C. Re-schedule with your friends so you can stay home and study. **(1)**
D. Go out with your friends, but leave early so you can come home to study. **(0)**

Facebook SJT (description-independent item; Schäpers, Lievens, et al., 2019)

You are on vacation with your friends. In your vacation home, you all use the same laptop. One day you browse your Facebook account but you forget to log out afterwards. Subsequently, your friends are having fun to comment and like posts, pictures, and Facebook pages on your behalf. What should you do?

A. You call up the tab "security" in the settings and check your activity on Facebook. If necessary, you undo the activities. **(0)**
B. You call up the Facebook activity log and check your latest activities on Facebook. **(1)**
C. You search your browsing history and your news feed for possible conspicuous features. **(0)**
D. Realizing that you are still logged in, you promptly call up your account on your laptop and log out again. **(0)**

Adapted Version of the Team Role Test (description-dependent item; Mumford et al., 2008; Schäpers, Mussel, et al., 2019)

You are the most experienced member of a newly formed production team with several members who are new to this type of manufacturing. The manufacturing process is complex, requiring compliance with precise standards, to avoid large amounts of product waste and possible equipment damage. Your supervisor has just informed your team that the sales department had made a "rush order", committing to ship a large batch of product five days before the anticipated ship date. What should you do?

A. Avoid being overly assertive in the new team and let others determine the teams direction, because it is important that the younger members take the lead. **(0)**
B. Quickly meet with your team members to decide the priority that should be given to the "rush order". **(1)**
C. Try not to react too strongly to the news to help the new team members understand that this kind of rush order occurs far too often. **(0)**
D. Suggest that the deadline is unreasonable, and you will simply have to do your best without worrying about meeting the unrealistic shipment date to which the Sales department committed themselves. **(0)**

Table S4

Itemwise Comparison of SJT Item Difficulties between Group 1 and Group 2 (Study 2)

Item	*d*	*t*	*df*	*p*
Description-independent				
TRT 1	.54	4.72	304.42	<.001
TRT 4	-.02	-.22	355	.586
PI 7	-.09	-.79	311.17	.786
PI 10	-.21	-1.73	281.06	.958
FB 4	.71	6.75	361.99	<.001
FB 8	.25	2.29	327.29	.011
Description-dependent				
TRT 3	.52	4.71	333.66	<.001
TRT 5	2.30	21.46	348	<.001
PI 1	.56	5.63	400	<.001
PI 9	.75	5.88	243.61	<.001
FB 1	.31	3.04	384	.001
FB 9	.12	1.09	384	.138

Notes. One-sided *t*-tests. Degrees of freedom vary due to the use of *t*-tests for homogeneous and heterogeneous variances. Higher effect sizes reflect more correct answers on items with situation descriptions compared with items without situation descriptions. TRT = Team Role Test, PI = Personal Initiative SJT, FB = Facebook SJT. Numbers of item names indicate the item positioning in the complete SJT. Sample sizes ranged from $n = 350 – 402$.

* $p < .00417$ (adjusted to account for alpha inflation).

Table S5

Multi-group Regression Analysis (Study 2; including three re-categorized SJT items; see Table S4)

	Comparison against Null Model			Comparison against Baseline Model			Relevant DIAMONDS			R^2		
	χ^2	*p*	Equality constraints	$\Delta\chi^2$	Δdf	*p*	G1	G2	G3	G1	G2	G3
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	group 1 & 2	5.767	3.92	.209	I, De, S	I, De, S	S	.25	.22	.14
PI 10	55.987	<.001	group 1 & 2	8.762	3.60	.052	D, I, S	D, I, S	O, N	.18	.20	.08
FB 8	51.704	<.001	group 1 & 2	5.061	2.96	.163	D, M	D, M	O, De	.14	.19	.12
FB 9	27.432	.285	all groups	7.153	5.41	.248	-	-	-	.03	.05	.03
Description-dependent items												
TRT 1	27.995	.260	all groups	5.599	6.40	.517	D, O	D, O	D, O	.09	.09	.12
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
FB 4	34.894	.070	all groups	7.966	5.69	.213	M, S	M, S	M, S	.07	.10	.06

Notes. 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. TRT = Team Role Test, PI = personal initiative SJT, FB = Facebook SJT, D = Duty, I = Intellect, A = Adversity, M = Mating, O = pOsitivity, N = Negativity, De = Deception, S = Sociality. Sample sizes ranged between $n_{group1} = 205 – 234$, $n_{group2} = 142 – 170$, $n_{group3} = 211 – 230$. * $p < .05$.

Figure S6. Interaction Plots for Situation Characteristics (Item x Group; Study 2). D = Duty, I = Intellect, A = Adversity, M = Mating, O = pOsitivity, N = Negativity, De = Deception, S = Sociality, TRT = Team Role Test, PI = Personal Initiative SJT, FB = Facebook SJT. Group 1 = complete SJT item, group 2 = only response options, group 3 = only situation description

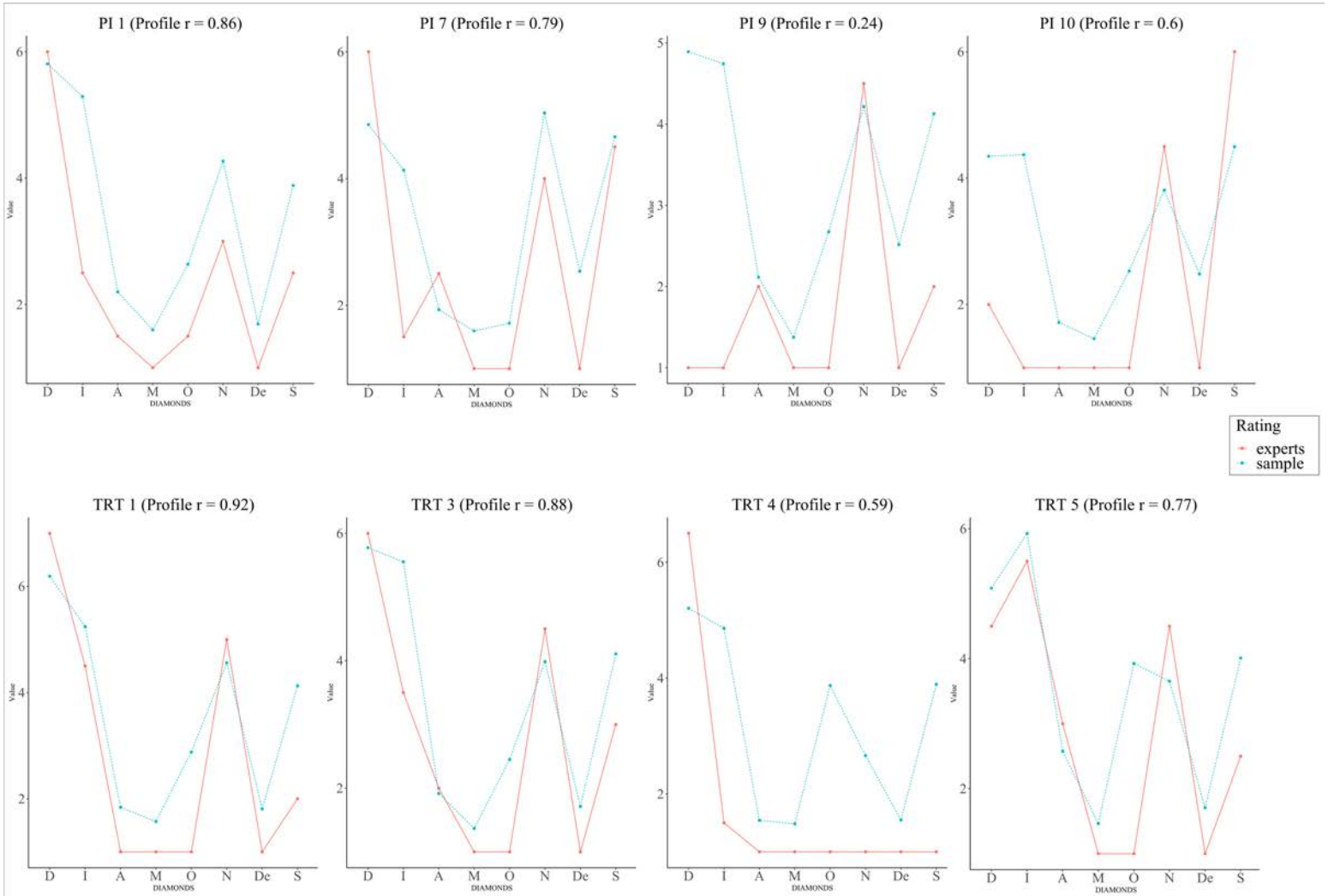


Figure S7. Comparison of Expert Ratings and Mean DIAMONDS of Group 1 (Study 2). D = Duty, I = Intellect, A = Adversity, M = Mating, O = pOsitivity, N = Negativity, De = Deception, S = Sociality, PI = Personal Initiative SJT, TRT = Team Role Test

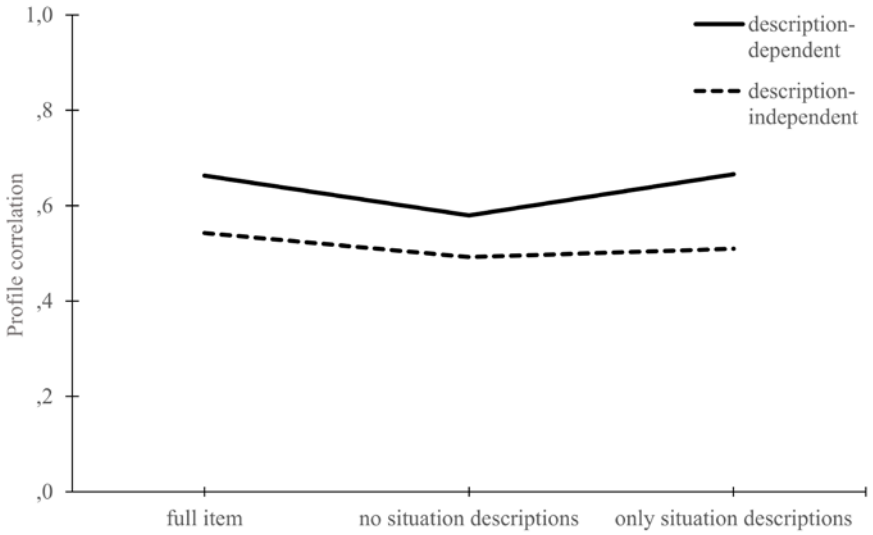


Figure S8. Interaction plot of two-way ANOVA (group x description-dependency) of mean profile correlation of individual's perception of DIAMONDS and correct situation construal as inferred from subject matter expert ratings (Study 2*).*

Table S8

Criterion-related Validity of Perceived Situation Characteristics when controlled for Personality, General Mental Ability, and Emotion Recognition (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	.096	S	.086	-	.054	I, S	.187
SJT PI 2	-	.088	-	.034	De	.040	-	.042
SJT PI 3	-	.021	M	.057	N	.048	-	.041
SJT PI 4	-	.029	-	.064	-	.035	-	.079
SJT PI 5	D, N, De	.220	N	.154	D, N, De	.215	-	.072
SJT PI 6	I	.123	D, A, M	.166	-	.117	N	.106
SJT PI 7	-	.115	M, N	.138	-	.026	-	.102
SJT PI 8	N, S	.122	A, N, De	.223	-	.064	A	.056
SJT PI 9	I	.066	-	.052	-	.044	-	.059
SJT PI 10	D, A, De	.124	D, I A, S	.177	D	.104	I	.045
SJT PI 11	I	.072	-	.075	-	.024	De	.124
SJT PI 12	A, De	.103	A, De	.122	De	.061	D, M	.136
SJT SC 1	-	.038	-	.106	-	.036	-	.037
SJT SC 2	-	.065	S	.125	-	.043	-	.067
SJT SC 3	-	.018	-	.048	-	.026	-	.064
SJT SC 4	N	.090	N	.126	N	.079	N	.148
SJT SC 5	-	.065	A	.166	O	.126	-	.053
SJT SC 6	M, O	.133	O, N	.160	O	.084	-	.097

Notes. DIAMONDS dimensions depicted refer to regression weights with $p < .05$. ΔR^2 refers to incremental explained variance of perceived situation characteristics in criteria above and beyond SJT performance, general mental ability, Big Five personality, and self-report personal initiative. PI = personal initiative, SC = self-consciousness, OCBI = organizational citizenship behavior, IRB = in-role behavior, D = Duty, I = Intellect, A = Adversity, M = Mating, O = pOsitivity, N = Negativity, De = Deception, S = Sociality. N = 285.

Table S10
Criterion-related Validity (Self-rated) of Perceived Situation Characteristics (Study 3)

	OCBI		IRB	
SJT items	DIAMONDS	ΔR^2	DIAMONDS	ΔR^2
PI 1	D, M	.057	D, M	.070
PI 2	I, O, S	.100	S	.053
PI 3	D	.048	D, A, M	.101
PI 4	O	.053	N	.042
PI 5	-	.017	-	.017
PI 6	D, I, O	.098	A, De	.081
PI 7	I, O	.100	-	.053
PI 8	-	.048	I, S	.056
PI 9	-	.031	D	.038
PI 10	M	.042	O	.043
PI 11	D, De	.081	D, O	.094
PI 12	D, M	.076	D, M	.057
SC 1	M, S	.080	S	.053
SC 2	I, O	.146	I, O	.091
SC 3	I, O	.110	I	.053
SC 4	N	.086	-	.032
SC 5	I, A, M	.090	A	.046
SC 6	A	.095	S	.078

Notes. 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. PI = personal initiative, SC = self-consciousness, OCBI = organizational citizenship behavior, IRB = in-role behavior, D = Duty, I = Intellect, A = Adversity, M = Mating, O = pOsitivity, N = Negativity, De = Deception, S = Sociality. $N = 285$.

References

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

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

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics, 46*, 1171-1178. https://doi.org/10.2307/2532457

Christensen, R. H. B. (2018). *Cumulative link models for ordinal regression with the R package ordinal*. Manuscript submitted for publication.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 425-438. https://doi.org/10.1080/10705511.2014.915373

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8*, 430-457. https://doi.org/10.1207/S15328007sem0803_5

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

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

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.

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

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

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

McFadden, D. (1979). Quantitative methods for analysing travel behaviour of individuals: Some recent developments. In D. A. Hensher & P. R. Stopher (Eds.), *Behavioural travel modelling* (pp. 279-318). London, UK: Croom Helm.

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

Moore, C. T. (2016). gtheory: Apply Generalizability Theory with R [Computer Software]. https://cran.r-project.org/web/packages/gtheory/index.html

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

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

Peters, G. (2018). *Userfriendlyscience: Quantitative analysis made accessible* [Computer software]. https://doi.org/10.17605/osf.io/txequ

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

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

Revelle, W. (2018). *Psych: Procedures for Personality and Psychological Research* [Computer software]. Evanston, IL: Northwestern University. https://cran.r-project.org/pacakage=psych

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

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?* 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

Sherman, R. A. (2015). *multicon: An R package for the analysis of multivariate constructs* (R package version 1.6) [Computer software]. Retrieved from https://CRAN.R-project.org/package=multicon

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

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

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

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.2013.12.00

Appendix E

Individual Contributions to Research Papers

Chapter 2: On the construct-related validity of implicit trait policies

Jan-Philipp Freudenstein:	Conceptualization, study designs, methodology, data collection, data analyses, writing, review and editing
Patrick Mussel:	Conceptualization, study design (Study 2), review and editing
Stefan Krumm:	Conceptualization, review and editing, supervision

Stephan J. Motowidlo and Filip Lievens provided helpful comments on an earlier version of this manuscript.

Chapter 3: Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance

Jan-Philipp Freudenstein:	Conceptualization, study designs, methodology, data collection, data analyses, writing, review and editing, project management, administration of publication process
Philipp Schäpers:	Conceptualization, study design (Study 1), review and editing
Lena Roemer:	Study design (Study 2), data collection (Study 2), review and editing
Patrick Mussel:	Conceptualization, review and editing
Stefan Krumm:	Conceptualization, study designs (Studies 2 & 3), review and editing, supervision

Alexandra Göbel contributed to the data collection as research assistant in the working group for psychological assessment, differential and personality psychology, Freie Universität Berlin (Prof. Dr. Stefan Krumm).

Cornelius König provided helpful comments on an earlier version of this manuscript.

The following theses contributed to data collection:

Breda, M. (2017). *How can the criterion-related validity of situational judgment tests (SJTs) be explained? Research on the influence of situation perception on SJT performance* (Unpublished bachelor's thesis). Freie Universität Berlin, Germany.

Harst, E. K. (2017) *Explorative Study on the relation between the situational eight DIAMONDS and situational judgment tests* (Unpublished bachelor's thesis). Freie Universität Berlin, Germany.

Haas, L. (2018). *Die acht DIAMONDS und berufliche Leistung: Der Einfluss von Situationswahrnehmung auf die Kriteriumsvalidität von Situational Judgment Tests* (Unpublished bachelor's thesis). Freie Universität Berlin, Germany.

Kalusa, J. (2019). *Wie wichtig ist die Situation in Situational Judgment Tests? Eine Untersuchung zum Einfluss der Situationswahrnehmung auf SJTs unter Kontrolle der Persönlichkeit* (Unpublished master's thesis). Psychologische Hochschule Berlin, Germany.

Kindermann, J. (2018). *Welchen Einfluss hat die Stärke von Situationen bei der Bearbeitung von situational judgment tests (SJTs)* (Unpublished bachelor's thesis). Freie Universität Berlin, Germany.

Klimitz, L. (2019). *Interindividuelle Unterschiede in der Verarbeitung situativer Reize: Die Bedeutung von Emotionserkennungsfähigkeit für die Situationswahrnehmung von Situational Judgment Tests* (Unpublished bachelor's thesis). Freie Universität Berlin, Germany.

Roemer, L. (2017). *To have a feel of situations – Is the perception of psychological situation characteristics related to contextuality in situational judgment test items?* (Unpublished master's thesis). Freie Universität Berlin, Germany.

Chapter 4: The influence of situational strength on the relation of personality and situational judgment test performance

Jan-Philipp Freudenstein:	Conceptualization, study design, data collection, data analyses, writing, review and editing
Philipp Schäpers:	Conceptualization, data collection, review and editing
Nomi Reznik:	Conceptualization, review and editing
Stefan Krumm:	Conceptualization, review and editing, supervision

The following bachelor thesis tested different methods to operationalize situational strength of SJT items. This study contributed to the decision on the final study design:

Kindermann, J. (2018). *Welchen Einfluss hat die Stärke von Situationen bei der Bearbeitung von situational judgment tests (SJTs)* (Unpublished bachelor's thesis). Freie Universität Berlin, Germany.

Chapter 5: Standardized state assessment: A methodological framework to assess person-situation processes in hypothetical situations

Jan-Philipp Freudenstein:	Conceptualization, development of framework, literature review, writing, review and editing
Julian Schulze:	Conceptualization, literature review, review and editing
Philipp Schäpers:	Conceptualization, review and editing
Patrick Mussel:	Conceptualization
Stefan Krumm:	Conceptualization, review and editing, Supervision

Nico Remmert contributed to the literature review as research assistant in the working group for psychological assessment, differential and personality psychology, Freie Universität Berlin (Prof. Dr. Stefan Krumm).

Appendix A: Developing a short-form situational judgment test to assess implicit trait policies for agreeableness

Jan-Philipp Freudenstein:	Conceptualization, study design, methodology, data collection, data analyses, writing, review and editing
Stefan Krumm:	Conceptualization, review and editing, supervision

Appendix B: English translation of the teamwork situational judgment test (SJT-TW*)*

Jan-Philipp Freudenstein:	Conceptualization, translation, data collection, data analyses, writing, review and editing
Nico Remmert:	Translation, writing, review and editing
Nomi Reznik:	Translation, review and editing
Stefan Krumm:	Translation, review and editing, supervision

Appendix C: Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests

Philipp Schäpers:	Conceptualization, writing, review and editing, administration of publication process
Jan-Philipp Freudenstein:	Conceptualization, writing, data analyses, review and editing
Patrick Mussel:	Conceptualization, study design, methodology, data collection, review and editing
Filip Lievens:	Conceptualization, review and editing
Stefan Krumm:	Conceptualization, study design, review and editing, supervision

Appendix F

Curriculum Vitae

Curriculum Vitae

Jan-Philipp Freudenstein (né Schulz)
Freie Universität Berlin
Habelschwerdter Allee 45
14195 Berlin, Germany

Education

10/2013 – 09/2016	Psychology (MSc), Ulm University
03/2015 – 07/2015	Semester abroad, Universidad de Chile
10/2012 – 09/2013	Business Psychology (BSc), Leuphana University of Lüneburg

Academic Experience

Since 10/2016	Research Associate, Psychological Assessment, Differential, and Personality Psychology, Freie Universität Berlin
06/2016 – 08/2016	Research Intern, Individual Differences and Psychological Assessment, Ulm University
02/2014 – 03/2016	Research Assistant, Work and Organizational Psychology, Ulm University
11/2015 – 03/2016	Research Assistant, Sustainable Management, Ulm University

Teaching Experience

Psychological Assessment	Courses on personnel selection and test theory analyses in R
Personality Psychology	Course on personality development and group differences in personality

Nine supervised bachelor theses
Three supervised master theses

Academic Self-Governance

Since 04/2019	Member of the Board of Examiners (Bachelor psychology), Freie Universität Berlin
Since 04/2019	Member of the Education Commission, Department of Education and Psychology, Freie Universität Berlin
Since 04/2019	Deputy member of Structural Commission for Psychology, Freie Universität Berlin
06/2014 – 09/2016	Head of Sustainability, Student's Union Executive Committee, Ulm University
10/2016 – 09/2016	Member of Student Body, Institute of Psychology and Education, Ulm University

Ad Hoc Reviewer

Organizational Psychology Review
Psychological Test Adaptation and Development

Professional Memberships

German Psychological Society (DGPs)
- Section for Personality Psychology, and Psychological Diagnostics
- Section for Industrial and Organizational Psychology

Grants

2019 Research grant, Department of Education and Psychology, Freie Universität Berlin *(1953€)*

2018 Travel grant, Department of Education and Psychology, Freie Universität Berlin *(216,78€)*

2017 Research grant, Department of Education and Psychology, Freie Universität Berlin *(2000€)*
Travel grant, German Psychological Society, Section for Industrial and Organizational Psychology *(200€)*

Publications

Journal Articles

Freudenstein, J.-P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*. https://doi.org/10.1111/peps.12385

Schäpers, P., **Freudenstein**, J.-P., Mussel, P., Lievens, F., & Krumm, S. (2020). Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests. *Journal of Research in Personality*. https://doi.org/10.1016/j.jrp.2020.103963

Schäpers, P., Lievens, F., **Freudenstein**, J.-P., Hüffmeier, J., König, C. J., & Krumm, S. (2020). 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*, *93*(2), 472–494. https://doi.org/10.1111/joop.12297

Freudenstein, J.-P., Strauch, C., Mussel, P., & Ziegler, M. (2019). Four personality types may be neither robust nor exhaustive. *Nature Human Behaviour*, *3*(10), 1045–1046. https://doi.org/10.1038/s41562-019-0721-4

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

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*(5), 475–476. https://doi.org/10.1002/per.2119

Other Publications

Freudenstein, J.-P., & Krumm, S. (2020). *Developing a short-form situational judgment test to assess implicit trait policies for agreeableness*. https://doi.org/10.31219/osf.io/kax7n

Schäpers, P., **Freudenstein**, J.-P., & Krumm, S. (2020). Situational Judgment Tests. In M. A. Wirtz (Ed.), *Dorsch—Lexikon der Psychologie*. Hogrefe.

Talks and Presentations

Invited Presentations

Freudenstein, J.-P. (2019, January). Explaining climate change. *Network analyses as a tool to understanding pro-environmental behavior*. Paper presented at the colloquium of the Individual Differences and Psychological Assessment working group, Ulm University.

Organized Symposia

Freudenstein, J.-P., Reineboth, M., & Krumm, S. (2019, September). *Understanding situatiational judgment tests. New insights into underlying constructs and psychometric properties*. Symposium at the 15th DPPD conference, Dresden, Germany.

Freudenstein, J.-P., Schäpers, P., & Krumm, S. (2019, May). *A closer look at situational judgment tests: New developments and insights*. Symposium at the 19th European Association of Work and Organizational Psychology (EAWOP) Congress, Turin, Italy.

Conference Presentations (first-authorships only)

Freudenstein, J.-P., Mussel, P., & Krumm, S. (2019, September). *Measurement artifact or fundamental process of situational judgment tests? Assessing the construct validity of implicit trait policies*. Paper presented at the 15th DPPD conference, Dresden, Germany.

Freudenstein, J.-P., Schäpers, P., Roemer, L., Mussel, P., & Krumm, S. (2019, May). *Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance*. Paper presented at the 19th European Association of Work and Organizational Psychology (EAWOP) Congress, Turin, Italy.

Schulz, J.-P., Schäpers, P., Mussel, P., & Krumm, S. (2018, September). *Führt empirical option weighting zur Verbesserung der psychometrischen Eigenschaften von Situational Judgment Tests*. Forschungsvortrag auf dem 51. Kongress der Deutschen

Gesellschaft für Psychologie, Frankfurt a. M., Deutschland.

Schulz, J.-P., Schäpers, P., & Krumm, S. (2017, September). *Messen Situational Judgment Tests die Wahrnehmung situativer Informationen?* Forschungsvortrag auf der 10. Tagung der Fachgruppe Arbeits- und Organisationspsychologie, Dresden, Deutschland.

Schulz, J.-P., Mussel, P., & Krumm, S. (2017, September). *Situational Judgment Tests: Welchen Einfluss hat die Situation auf die Konstruktvalidität?* Forschungsvortrag auf der 14. Arbeitstagung der Fachgruppe Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik, München, Deutschland.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, den 22.06.2020

Jan-Philipp Freudenstein