

Fachbereich Erziehungswissenschaft und Psychologie
Freie Universität Berlin

**Situational by Structure: Rethinking the Role of Situation Construal in
Situational Judgment Tests**

Dissertation

Zur Erlangung des akademischen Grades

Doktorin der Philosophie (Dr. phil.)

Vorgelegt von

M.Sc.

Reznik, Nomi

Berlin, September 2025

Erstgutachter

Prof. Dr. Stefan Krumm (Freie Universität Berlin)

Zweitgutachter

Prof. Dr. Philipp Schäpers (Universität Münster)

Disputation: 17.12.2025

Table of Contents

Acknowledgements	i
Summary	ii
Zusammenfassung	iii
List of Tables	v
Chapter 1: Introduction	1
Chapter 2: ATIC in SJTs	27
Chapter 3: SJTs and Their Components	51
Chapter 4: SJTs and Their Situation Construal	108
Chapter 5: Discussion	154
Author Contributions	vi
Eidesstattliche Erklärung	vii

Acknowledgements

Writing a dissertation is, as I've come to know, a process that is maybe not impossible but definitely highly unlikely to finish without the support of many, many people, and I'd like to take this opportunity to thank them.

First, I want to thank Stefan Krumm, who trusted me to do this research, and to do it well. Thank you for giving me this opportunity in the first place, for our talks and feedback sessions, for fighting the battles of academic bureaucracy for and with me.

Thank you to Philipp Schäpers, for answering all my questions immediately, no matter the ungodly hour I asked them.

Thank you to all my current and former colleagues-turned-friends: thank you, Lida and Alyce, for our never-ending jokes, for ranting together, for coming up with solutions for each and every problem imaginable. Thank you, Nico and Franz, for being the best Nachwuchsförderungsnetzwerk anyone could ever ask for, for keeping connected and updated, and for sitting in my first real-life conference talk ever (DGPs 2022!) to show support, even when I said "don't come, it's going to be lousy" (it was okay). Thank you, Jan-Philipp, for your research (and life) advice, your help that I could always count on, and for all the memes.

Thank you to my family and friends: Thank you to Juffi for being on-call literally all the time. Thank you to Leon and Rolf for dealing with my dissertation rants for all these years and still supporting me to go on. Thank you to my parents for never stopping to explain that their daughter is, yes, a psychologist, but no, not a therapist. Thank you to Omi for reading every word I ever wrote (this one's a page-turner!). Thank you to Jana for reality checks every five minutes. Thank you to Sascha, for everything.

Summary

Situational judgment tests (SJTs) are widely used tools in personnel selection, yet their theoretical foundations remain debated. While traditionally described as contextualized simulations that capture dynamic person-situation processes, recent work has raised doubts about whether SJTs actually require situation construal or rather rely on structural item properties. In my dissertation, I develop and test a working model of SJT responding that integrates person factors (e.g., trait social desirability), item components (e.g., situation descriptions, response options, trait-relevant cues), and the psychological interpretation of the situation (situation construal via ATIC, DIAMONDS, situational strength, and social desirability perceptions). First, I examine ATIC as an individual difference in inferring evaluation criteria, extending its application to SJTs and testing its predictive power under varying incentive conditions. Second, I manipulate SJT item versions to isolate situation descriptions and response options, showing that response options carry most of the construct and criterion validity. Third, I directly test whether construal ratings predict performance across item versions, with results pointing to weak and inconsistent links. Together, these findings suggest that the “situationality” of SJTs lies less in dynamic construal processes and more in their structural design, particularly the information embedded in response options. The dissertation contributes to theory by reframing SJTs as structurally contextualized assessments and to practice by highlighting design properties such as cue quality and construct alignment.

Keywords: situational judgment tests, situation construal, ATIC, trait activation theory, social desirability, DIAMONDS, construct validity, criterion validity

Zusammenfassung

Situational Judgment Tests (SJTs) sind weit verbreitete Instrumente bei der Personalauswahl, doch ihre theoretischen Grundlagen sind nach wie vor umstritten. Während sie traditionell als kontextbezogene Simulationen beschrieben werden, die eine dynamische Interaktion zwischen Person und Situation erfassen, haben neuere Arbeiten Zweifel daran aufkommen lassen, ob SJTs tatsächlich situation construal benötigen, um beantwortet zu werden, oder ob sie eher auf strukturellen Eigenschaften der Items basieren. In meiner Dissertation entwickle und teste ich ein Arbeitsmodell für die Beantwortung von SJTs, das Personenmerkmale (z. B. soziale Erwünschtheitsdisposition), Item-Komponenten (z. B. Situationsbeschreibungen, Antwortoptionen, trait-relevant cues) und die psychologische Interpretation der Situation (situation construal per ATIC, DIAMONDS, Situationsstärke und Wahrnehmung sozialer Erwünschtheit) integriert. Zunächst untersuche ich ATIC als interindividuellen Unterschied in der Erkennungsfähigkeit von Bewertungskriterien, erweitere dessen Anwendung auf SJTs und teste dessen Vorhersagekraft unter verschiedenen Anreizbedingungen. Zweitens manipulierte ich SJT-Item-Versionen, um Situationsbeschreibungen und Antwortoptionen zu isolieren, und zeige, dass Antwortoptionen den größten Teil der Konstrukt- und Kriteriumsvalidität bewirken. Drittens teste ich direkt, ob construal-Bewertungen die Leistung über Item-Versionen hinweg vorhersagen, wobei die Ergebnisse auf schwache und inkonsistente Zusammenhänge hinweisen. Zusammengefasst deuten die Ergebnisse darauf hin, dass die „Situationsabhängigkeit“ von SJTs weniger in dynamischen construal-Prozessen als vielmehr in ihrem strukturellen Design liegt, insbesondere in den in den Antwortoptionen enthaltenen Informationen. Die Dissertation leistet einen theoretischen Beitrag, in dem sie SJTs als strukturell kontextualisierte Instrumente neu definiert, und einen praktischen Beitrag, in dem sie Designbestandteile wie cue-Qualität und Konstruktpassung hervorhebt.

Keywords: Situational Judgment Tests, situation construal, ATIC, Trait Activation Theory, soziale Erwünschtheit, DIAMONDS, Konstruktvalidität, Kriteriumsvalidität

List of Tables

Chapter 2: Main text

Table 2.1 Means, SDs, and Bivariate Correlations between ATIC and SJT Item Scores	33
Table 2.2 Results of Mixed-Effects Model for ATIC Predicting SJT Responses.....	33
Table 2.3 Means, SDs, and Bivariate Correlations between ATIC and SJT Item Scores (Low-Incentive)	36
Table 2.4 Means, SDs, and Bivariate Correlations between ATIC and SJT Item Scores (High-Incentive)	36

Chapter 2: Appendices

Table 2.B1 Items per Groups and Items Excluded from the Initial Item Pool	44
Table 2.C1 Means, SDs, and Bivariate Correlations between ATIC and SJT Item Scores (A-Set)	46
Table 2.C2 Means, SDs, and Bivariate Correlations between ATIC and SJT Item Scores (B-Set)	47
Table 2.C3 Means, SDs, and Bivariate Correlations between ATIC and SJT Item Scores (C-Set)	49

Chapter 3: Main text

Table 3.1 Descriptive Item Statistics and Scale Reliability for SJT Item Versions i, iii, and iv	67
Table 3.2 Descriptive Item Statistics and Scale Reliability for SJT Item Version ii	68
Table 3.3 Descriptive Statistics and Bivariate Correlations for Self- and Supervisor Reports with SJT Items	69
Table 3.4 Hierarchical Regression Analyses with SJT Item Versions i–iii Predicting Item Version iv	71
Table 3.5 Incremental Contribution of Version ii in Predicting Self-Reports	74
Table 3.6 Moderated Regression Analyses with Trait Social Desirability (TSD)	75

Chapter 3: Appendices

Table 3.B1 Welch’s t-Test Results for Sequence Effects (Different Wave Sequence)	97
Table 3.B2 Welch’s t-Test Results for Learning Effects (Previous Wave Exposure)	99
Table 3.C1 Relative Weights for SJT Versions I to iii in Predicting the Full Item Version	100
Table 3.D1a Hierarchical Regression Analyses (Personal Initiative; self & supervisor reports)	101
Table 3.D1b Hierarchical Regression Analyses (Personal Initiative; reversed steps)	103
Table 3.D2a Hierarchical Regression Analyses (Conscientiousness; self & supervisor reports)	105
Table 3.D2b Hierarchical Regression Analyses (Conscientiousness; reversed steps)	106

Chapter 4: Main text

Table 4.1 Regression Analyses Predicting Item Performance in the Full Item	130
Table 4.2 Interaction Effects for DIAMONDS between Item Variations Containing vs. Not Containing Cues	131
Table 4.3 Mixed-Effects Regression for Situational Strength and Social Desirability	133

Chapter 4: Appendices

Table 4.A1 Measurement Invariance for Situational Strength	150
Table 4.B1 Mixed-Effects Regression for Situational Strength and Social Desirability (Full Model)	151

Chapter 1

Introduction

Introduction

There is a rather well known philosophical question that goes “If a tree falls in the forest and there is no one there to hear it, does it make a sound?”. While this thought experiment has been attributed to multiple famous philosophers, it is actually unknown who first came up with it. Still, it can be adapted to fit a wide variety of topics, for example: if a situation takes place, but no one experiences it as a situation, was it even a situation after all? Or, in a more psychological than philosophical framing: what actually *is* a situation? Among the numerous questions in the field of psychology that researchers have asked for decades now, striving to understand and explain human behavior and experience, this one was asked pretty early on (Lewin, 1936). This is rather unsurprising, as humans behave and experience everything *within* a situation of some sorts, there is no human experience within complete informational vacuum, as our senses construct something (i.e., a *situation*) from everything surrounding us. It is a common, everyday observation, that two people physically present in the same location at the same time might have very differing perceptions, behaviors, and recollections of the (deceptively worded) same event. Lewin equated human behavior to the function of person and environment as early as 1936, and was onto something: his equation $B = f(P, E)$ provided the basis for further theoretical and empirical research on how individuals’ psychological construction of the world around them works. Famously, Mischel (1977) theorized how different the effects of objective stimuli can be, depending on how any given individual “construes and transforms” (p.253), that is, *perceives*, information around them. From there on, the field of person-situation-interaction took off (Bem & Allen, 1974; Funder, 2006, 2016; Funder & Colvin, 1991; Mischel & Shoda, 1995). Every behavior or experience assessed by any type of psychological assessment can be understood to take place within the situation-construal model (SCM) proposed by Funder (2016): within the SCM, behavior is influenced

Chapter 1: Introduction

not only by dispositional traits of the individual and the objective traits of the environment, but specifically by how each individual *construes* the situation. This construal process is shaped by both the objective characteristics of the situation (i.e., societal rules, incentives, norms...) and the personality of the individual, but remains an additional determinant of behavior. The SCM builds on Lewin's (1936) theory, Mischel's (1977) emphasis on perception, and more recent work by Rauthmann et al. (2014) on situational taxonomies.

This understanding of human behavior as the result of individual traits, objective situation traits, the subjective situation construal and their interaction has profound implications for applied domains of psychology; specifically for personnel selection, where human behavior takes place in very specific, structured settings, and most prominently in simulation-based assessments. Assessment Centers (ACs), structured interviews, and situational judgment tests (SJTs) are all designed to elicit job-relevant behaviors by placing candidates in standardized yet realistic scenarios, with the idea that individuals behave similarly across comparable situations (Tett & Guterman, 2000; Wernimont & Campbell, 1968). The foundational assumption is that these scenarios function as psychologically meaningful situations, thereby allowing for observable behaviors that reflect underlying traits or abilities (Jansen et al., 2013). These traits are, according to Trait Activation Theory, expressed only when individuals *perceive*, that is, construe, relevant cues (Tett & Burnett, 2003) from the situation, and research has shown that the ability to not only perceive but also explicitly identify these situational demands (the Ability To Identify Criteria, ATIC; Kleinmann, 1993) is linked to better performance both in the simulation and subsequent real-life work situations (Jansen et al., 2010; König et al., 2007). In this sense, situation perception becomes a prerequisite for valid assessment in personnel selection. While ACs and interviews often rely on overt behavioral tasks or interpersonal interaction, SJTs present scenarios in a more standardized, and less open format: Typically, they consist of written or video-based vignettes, i.e. descriptions of a situation, followed by several

response options (Motowidlo et al., 1990). Despite the conceptualization of SJTs as less situationally rich than ACs, they are similarly intended to simulate job-relevant situations, relying on the same core assumption: that test-takers will perceive these scenarios as psychologically meaningful and respond in ways that reflect their dispositional tendencies (Lievens & Sackett, 2012). A decade ago, this core assumption of SJT functioning was challenged, as Krumm et al. (2015) were able to show that SJTs could largely be still solved correctly when the situation description was omitted. This raises an important question, which will be the focal point of this dissertation: to what extent are SJTs actually truly “situational”?

How Situational Are Situational Judgment Tests?

Situational judgment tests (SJTs) have been around for quite a while, with the How Supervise?-instrument being arguably one of the earliest examples (File, 1945). However, they only reached real popularity in personnel selection about half a century later (Weekley & Jones, 1999). Originally, they were conceptualized as so-called low-fidelity simulations (Motowidlo et al., 1990) that were thought to function alike assessment centers, albeit cheaper and easier to administer (Lievens et al., 2003; Weekley et al., 2015): by simulating a situation similar to those that arise in real-life job situations, incorporating relevant trait activating information, and capturing test-takers’ subsequent behavior. Building on behavioral consistency, i.e. that individuals show similar behavior in similar situations (Tett & Guterman, 2000), behavior in the simulated situation would directly predict behavior in a similar, real-life scenario. For SJTs, ample meta-analytical evidence has since supported this idea as criterion-related validity of SJTs (that is, prediction of job performance) remains solid, especially in relation to their cost-effectiveness and low participant burden (Christian et al., 2010; McDaniel et al., 2007; McDaniel & Nguyen, 2001; Webster et al., 2020). At the same time, SJTs’ construct-related validity was shown to be consistently lacking (McDaniel et al., 2016), and another yet

Chapter 1: Introduction

unexplained problematic fact about SJTs was revealed: Early SJT research assumed that the situation descriptions would drive test performance by eliciting response choice to job-relevant scenarios (Motowidlo et al., 1990), just as performance in assessment centers is assumed to result from how candidates perceive and respond to situational demands (Jansen et al., 2013). A study by Krumm et al. (2015), however, empirically challenged this prevailing assumption: they found that across three studies, between 43 and 71% of examined SJT items could be solved correctly without the situation description, and strengthened the idea that SJTs might not behaviorally measure specific traits, but a broader, generalized domain knowledge (Motowidlo et al., 2009).

This new insight was followed by several studies, both empirical and theoretical, with similar ideas, producing similar results: Researchers systematically examined whether it is the response options in SJTs, rather than the situational descriptions, that may themselves function as situational stimuli, i.e. contain trait-relevant cues, thereby driving performance. Leeds (2018), for example, approached the question from a psychophysical perspective, proposing the Theory of Cognitive Acuity, which holds that SJT performance hinges on test-takers' ability to detect subtle differences in the effectiveness signals of response options. In this model, judgment accuracy depends less on situation construal and more on sensitivity to contrasts between response options. In a similar vein, Kaminski et al. (2019) demonstrated that SJT item response options carry substantial social desirability and plausibility cues, which explained a significant portion of response variance even when the associated situation descriptions were removed.

Schäpers et al. (2020) found that removing situation descriptions from construct-driven SJTs had no detrimental impact on construct-related validity, and in some cases (e.g., for items measuring conscientiousness), it was even enhanced. These findings in part echo a long-standing concern in SJT research: that their construct-related validity is so inconsistent, it has

been called a proverbial “hot mess” (McDaniel et al., 2016). When performance can be preserved or even improved without the situation descriptions, this challenges the assumption that SJTs measure traits via behavioral simulations. Instead, such findings suggest that test-takers may rely on other response strategies, blurring the link between intended item content and target constructs. A possible example for these other strategies was brought forth by Melchers and Kleinmann (2016) in response to the findings of Krumm et al. (2015), who argued that it is still situational information that influences response choice in SJTs, but that trait-relevant cues embedded in response options may be enough to activate personality expressions independently of situational context, i.e. that test-takers construe the situation from the information within the response options. However, this represents a shift in how SJTs function cognitively. In the traditional sense of simulation-based SJTs, the situation description provides a coherent, job-relevant context or a psychologically meaningful scenario to which the test-taker responds (Motowidlo et al., 1997; Motowidlo & Tippins, 1993). In contrast, when situational information is inferred from response options alone, this structure is lost: the test-taker is no longer placed into a simulated work situation but is instead reconstructing one based on behavioral information from the response options. While both pathways may lead to trait activation, the latter lacks the contextualization and psychological framing that traditionally define SJTs as simulations. This distinction is central to understanding what (and how) SJTs measure.

Freudenstein et al. (2020) added to this shift in perspective by arguing that SJT performance may rely more on the situational strength and construal of response options than on rich, scenario-based context. An alternative direction was pursued by Rockstuhl et al. (2015) who did not discard situational context but instead proposed that test performance could be better understood by explicitly measuring how accurately test-takers perceive and interpret the core situational judgment themes embedded in SJT items. They differentiated between situation

Chapter 1: Introduction

judgment accuracy, i.e. the ability to correctly identify the underlying situational theme, and traditional SJT response accuracy, and found situation judgment accuracy accounted for incremental variance in external criteria, above and beyond of what was explained by response judgment (that is, traditional SJT response accuracy).

Cumulatively, this line of research provides stable and consistent evidence that the presence of situation descriptions is not a necessary condition for eliciting correct responses in SJTs (Krumm et al., 2015) and in some cases, their removal may even enhance construct-related validity (Schäpers et al., 2020). These findings converge on the interpretation that response options themselves may function as trait-relevant informational cues, sufficient to activate situation construal, personality expressions or decision-making processes in the absence of explicit contextual framing (Freudenstein, Schäpers, et al., 2020; Kaminski et al., 2019; Leeds, 2018; Melchers & Kleinmann, 2016). Simultaneously, other studies emphasize that construal based on situation descriptions can make a unique, incremental contribution to predicting performance criteria (Rockstuhl et al., 2015), suggesting that criterion-related validity may, at least in part, be driven by the ability to accurately interpret the situational demands embedded in the item stem, rather than solely through response options. These dual perspectives indicate that the functional situational components of SJTs may be distributed across item components, rather than localized solely in the item stem or solely within the response options. These findings lead to a more precise refinement of the previously asked question regarding SJTs' situational content: How do test-takers actually construe SJT items as psychologically meaningful situations?

The Working Model of SJT Responding

To answer this question, detailed information about SJT response processes is needed. Previously, researchers have theorized three main SJT process models (the situated reasoning

Chapter 1: Introduction

and judgment model (SiRJ), Grand, 2020; the Tripartite Model, Martin-Raugh & Kell, 2021; and the predictor response process model, Ployhart, 2006). Although differing in terminology, all three models describe a similar cascade of SJT item understanding and interpreting, and finally choosing a response based on this interpretation. Ployhart (2006) calls this comprehension, where individuals interpret the situation and its cues, followed by retrieval, in which relevant prior knowledge is accessed to understand the situational demands. These steps inform judgment (evaluating the situation) and culminate in response selection. Grand's (2020) SiRJ model proposes a similar sequence: conditional reasoning, where test-takers interpret the item's contextual demands; similarity judgment, where response options are evaluated in relation to prior experience; and preference accumulation, in which the final choice is made based on the perceived appropriateness of the responses. Likewise, the Tripartite model by Martin-Raugh and Kell (2021) starts with situation perception, where objective information is transformed into psychologically meaningful features. This leads to goal formulation, in which the test-taker defines their intent in relation to the situation, and finally response evaluation, where response options are judged based on their utility for achieving the established goal. All three models converge on the idea that interpreting the situation is a fundamental cognitive mechanism driving behavior in SJTs. They form the conceptual foundation of this dissertation's working model of SJT responding, which integrates these models and assumes that situation construal precedes and informs response choice. To specify what specifically constitutes situation construal in the context of SJTs in this dissertation, the following section introduces several conceptual operationalizations that inform how individuals combine situational information.

Situational Strength

The concept of situational strength (Meyer et al., 2010) posits that situations vary in the degree to which they constrain behavior. Strong situations (e.g., with high clarity, consistency,

Chapter 1: Introduction

constraints, and consequences) tend to suppress the influence of individual differences, while medium to weak situations allow for greater trait expression. Within the working model, situational strength is conceptualized as a subjective component of situation construal reflecting how clearly, consistently, etc. test-takers perceive the demands of a given SJT item. As such, it is treated in parallel to other construal-forming dimensions like DIAMONDS (Rauthmann et al., 2014) and perceived social desirability: all describe how a situation is psychologically represented by the individual, rather than what the situation objectively contains.

DIAMONDS Taxonomy of Situational Characteristics

The DIAMONDS framework (Rauthmann et al., 2014) was derived from the SCM and identifies eight psychologically meaningful dimensions along which individuals perceive and construe situations: Duty, Intellect, Adversity, Mating, Positivity, Negativity, Deception, and Sociality. These dimensions provide structured vocabulary for how individuals make sense of situations and are particularly relevant for understanding perception and construal processes in the first two phases of SJT responding. Within the working model, DIAMONDS dimensions function as perceptual anchors that shape whether and how an individual interprets a given situation as task-relevant, threatening, morally loaded, or otherwise psychologically relevant.

Ability to Identify Criteria (ATIC)

ATIC (Kleinmann, 1993) reflects test-takers' ability to infer the evaluative standard of a given assessment situation. Originally developed in the context of assessment centers, ATIC has been shown to predict performance in both ACs and situational interviews (Ingold et al., 2015; König et al., 2007). In the working model, ATIC is conceptualized as an element of situation construal: recognizing the measured criteria of an SJT item is part of construing the situation. This aligns with prior applications of ATIC in assessment centers and interviews, where accurately identifying the criterion is seen as central to performance. In the context of

Chapter 1: Introduction

SJTs, inferring what the test is “about” becomes part of forming a psychologically meaningful situation representation.

Social Desirability Perception

Social desirability refers to test-takers’ tendency to respond to assessment in a way that portrays them favorably (Wiggins, 1966). Kaminski et al. (2019, and similarly, Brown & Martin-Raugh, 2024) found social desirability to be a relevant driver of SJT responding. In the working model, perceived social desirability of SJT items constitutes a part of situation construal as the extent to which a situation requires (or enables) socially desirable behavior shapes test-takers’ mental representation of the situation and influences specifically the second and third phase of SJT responding by which response options are judged socially (un-) desirable.

Effectiveness

According to the Theory of Cognitive Acuity (Leeds, 2018), SJT performance reflects individuals’ ability to discriminate between differences in effectiveness across response options. Drawing from signal detection theory (Leeds, 2012), this view treats response selection as a sensitivity task, where effectiveness judgments drive SJT performance independently of full situational understanding. In the working model, effectiveness comes into play similarly to social desirability, mainly between the second and third phase, as effective behavior needs to be identified and response options are judged regarding their individual effectiveness.

Trait-relevant cues

According to trait activation theory (Tett & Guterman, 2000; Tett et al., 2021), traits are expressed behaviorally only when activated by relevant situational cues. In the context of SJTs, these cues must be embedded within the item, either in the situation description or the response options, or both. The working model therefore conceptualizes the presence (or lack thereof) of

Chapter 1: Introduction

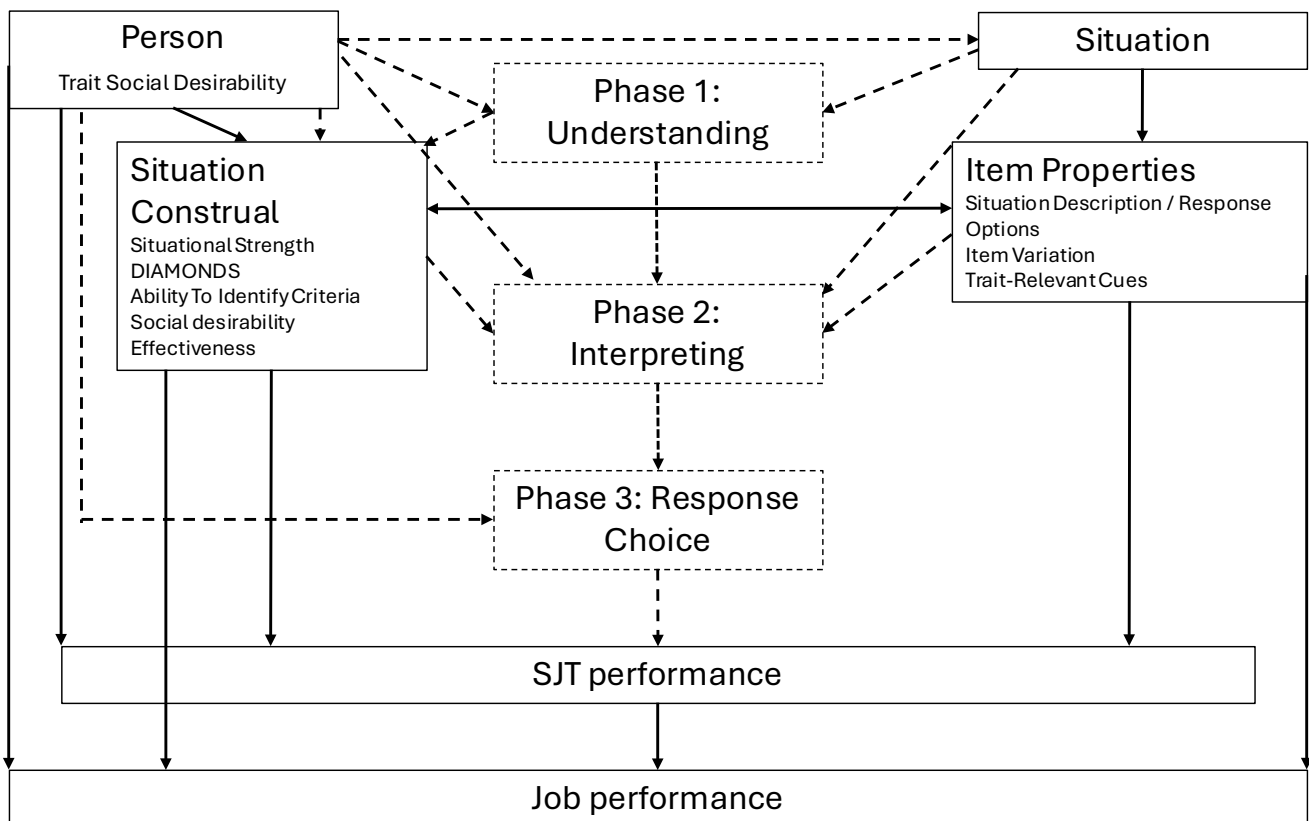
trait-relevant cues as a situational feature. However, for trait activation to occur, such cues must be recognized and understood by the test-taker, meaning that situation construal acts as the mechanism that determines whether available trait-relevant cues are cognitively processed and used in the three item response phases. This raises two critical questions for SJTs: not only whether trait-relevant cues are present, but also whether they are actually perceived as such by test-takers.

Building on this foundation, in the present dissertation I propose an integrative working model of SJT response processes (see Figure 1) structured into three sequential cognitive phases: understanding, interpreting, and response choice. These phases reflect the shared logic of the three process models, while integrating key components from situation perception theory and applied personnel selection research. In the understanding phase, test-takers process the presented item in the presented form, be it including both the situation description and the response options, one of them, or variations thereof, and begin forming an initial representation of the given situation. In the interpreting phase, this representation is shaped into a psychologically meaningful construal, influenced by both item characteristics (e.g., trait-relevant cues, variation of item presentation), person variables (such as goals or personality traits, which are not a central part of this dissertation), and individual situational construal variables. This phase reflects the cognitive operations underlying situation construal, where test-takers integrate situational information regarding situational strength, characteristics (i.e., the DIAMONDS), the construct measured by the item (i.e., ATIC), and the social desirability and effectiveness of behaviors facilitated by the item. Finally, in the response choice phase, construal then informs the selection of a response option, which then results in SJT performance. This performance may, in turn, predict external criteria like job performance.

The model presented here is a working conceptual framework, developed to clarify how I assume the cognitive phases of SJT responding based on established process models may align with the assumptions of situation construal (Funder, 2016). It is not intended as a formal structural model to be tested statistically but serves as a visual synthesis that illustrates which aspects of SJT item processing (e.g., understanding, interpreting, response choice) are thought to be influenced by different components of situation construal. Accordingly, the arrows and links in the model should be interpreted as conceptual mappings, not as hypothesized causal relationships.

Figure 1

Working Model of SJT Responding



Note. Visual representation of model integration across three SJT process models incorporating situation construal.

As delineated above, empirical evidence about the drivers of SJT responding is plentiful, but not exhaustive at all. Prior research has shown that there is *something* about the response

Chapter 1: Introduction

options alone, and there is also *something* about the situation descriptions alone, and there is also *something* about the whole item including both. Some overarching questions remain: how situational are SJTs after all, and which part of them is actually responsible for situationality, item performance, and criteria prediction? This dissertation addresses these questions through four empirical studies presented in three papers. The following chapters detail these studies, each contributing a different perspective on how test-takers construe SJTs and which factors influence their performance.

Chapter 2: Situational Judgment Tests and the Ability to Identify Criteria (ATIC)

ATIC is a cognitive construct originally introduced to explain why individuals vary in their performance in simulation-based selection methods such as assessment centers and structured interviews (Kleinmann, 1993). It describes the extent to which individuals are able to understand and explicitly identify what construct is being measured in a given task. In non-transparent, complex assessment situations where instructions are limited and behavioral expectations are implicit, ATIC has consistently emerged as a predictor of performance (Ingold et al., 2015; Jansen et al., 2010; Kleinmann et al., 2011; König et al., 2007; Melchers et al., 2009). Given that SJTs traditionally are thought to function similarly, i.e. are thought to also involve implicit behavioral demands, ATIC has been proposed as a relevant determinant of SJT performance. SJTs require test-takers to interpret social or work-related scenarios and evaluate response options, often under behavioral but non-specific instructions (e.g., “What would you do?”). Theoretically, ATIC in SJTs should also be instrumental in understanding what the test is actually assessing (phase 1) and selecting responses accordingly (phase 3). In that sense, ATIC reflects a meta-cognitive part of situation construal, enabling test-takers to align their responses with the untransparent trait demands of the situation. Empirical evidence for this assumption, however, remains yet limited and inconsistent. While some studies have reported moderate correlations between ATIC and SJT performance (Melchers & Hupp, 2017; Oostrom

Chapter 1: Introduction

et al., 2016; Wang et al., 2023), others have failed to find robust effects (Wolcott et al., 2021). The studies used differing presentations of SJT items and ATIC measures, and the mixed results raise the question of whether ATIC functions similarly across SJT types, formats, and instructions. Generally, however, to assess ATIC, test-takers' post-test identification of the construct targeted by each item is scored for accuracy against expert ratings (Ingold et al., 2015; Kleinmann, 1993). Given these theoretical and empirical considerations, the present paper investigates the role of ATIC in SJT performance across two studies. The first study examines whether ATIC predicts performance across items from multiple SJTs targeting multiple constructs in a planned-missingness design. The second study extends this question by focusing on one whole SJT measuring teamwork (Freudenstein, Remmert, et al., 2020) and testing whether the relationship between ATIC and performance changes under different motivational instructions (high-stakes vs. low-stakes).

Chapter 3: Situational Judgment Tests and Their Components

Typically, SJTs consist of a situation description followed by multiple response options from which test-takers select or rank responses (Motowidlo et al., 1990). However, while SJTs are frequently assumed to function as holistic simulations of real-life job situations, debate continues over which specific components of an SJT (situation stems, response options, or both) are essential for construct and criterion validity (Freudenstein, Schäpers, et al., 2020; Kaminski et al., 2019; Schäpers et al., 2020; Schäpers, Lievens, et al., 2019; Schäpers, Mussel, et al., 2019). Building on this yet ongoing debate, recent research has proposed diverging process models of how test-takers interact with SJT items. Models such as the Predictor Response Process Model (Ployhart, 2006), the Situated Reasoning and Judgment Framework (Grand, 2020), and the Tripartite Model (Martin-Raugh & Kell, 2021) all suggest that situational understanding underlies or precedes response evaluation and selection, in that test-takers rely on situation descriptions to understand the item before choosing a response.

Chapter 1: Introduction

However, empirical work increasingly challenges this assumption (Krumm et al., 2015; Schäpers et al., 2020; Schäpers, Lievens, et al., 2019), as researchers found that removing situation descriptions did not consistently impair SJT performance, suggesting that the response options alone might be enough for test-takers to infer the situation from. Similarly, Jackson et al. (2017) demonstrated that individual (rather than situational) variance predominantly drives SJT responses, further complicating the picture. Additional evidence for the importance of response options comes from research showing that they can independently activate personality traits: Kaminski et al. (2019) showed that social desirability and plausibility ratings of SJT responses predicted participant responding, even when situation descriptions were removed. Leeds (2018) with the proposal of the Theory of Cognitive Acuity, which views SJT responses judging between different extents of effectiveness, highlighted the informational value of response options themselves. Other work has suggested that test-takers can actively construe situations from the full set of responses (Freudenstein, Schäpers, et al., 2020), while in contrast, others like Rockstuhl et al. (2015) defended the value of situation stems through open-response formats similar to situational interviews (Latham & Sue-Chan, 1999) for their superior criterion-related validity. Overall, while theoretical models emphasize the role of situation comprehension, empirical findings on the role of situation descriptions and response options are inconclusive. This ambiguity motivates the central aim of chapter 3: to deconstruct SJTs into their individual components and assess their relative and combined contributions to construct- and criterion-related validity. Specifically, the study evaluates four item formats (response options only, randomized response options under different instructions, open-ended responses to situation descriptions, and full items) to shed light on which parts of SJTs drive performance, and construct- and criterion-related validities.

Chapter 1: Introduction

Chapter 4: Situational Judgment Test Components and Their Individual Situation Construal Properties

Do SJTs in fact trigger the kind of situation construal processes theorized to precede behavior in the SCM (Funder, 2016)? Building on person-situation interactionist theories (Funder & Colvin, 1991; Mischel, 1968) chapter 4 conceptualizes situation construal as a mechanism in which individuals transform objective information into psychologically meaningful characteristics (phase 1-2 in the working model). These characteristics then form the foundation for behavioral response selection (phase 3). Trait Activation Theory (Tett & Burnett, 2003) suggests that behavioral responses depend on the presence and interpretation of trait-relevant cues, which may be embedded in various parts of SJT items. Notably, the extent of trait expression hinges on moderate situational strength, as overly strong situations reduce the behavioral expression of traits (Harris et al., 2016; Marshall & Brown, 2006). Given these foundations, chapter 4 examines whether situation construal predicts performance in full SJT items and whether this relationship is moderated by the presence of trait-relevant cues. A further aim was to investigate whether construal is localized more in the item stem (situation description) or in the response options. Based on prior research (Freudenstein et al., 2020; Schäpers et al., 2020), both item parts may contain situational information, but their respective contributions remain unclear.

To capture situation construal empirically, the study drew on multiple established frameworks. First, the DIAMONDS taxonomy (Rauthmann & Sherman, 2015) was used to assess perceived situational characteristics. Second, situational strength (Meyer et al., 2014) was included as a construct reflecting how clearly a situation can be understood in describing and eliciting potential behavior. Third, perceived social desirability was assessed, conceptualized as a psychologically constraining factor akin to situational strength (Brown & Martin-Raugh, 2024; Golubovich, 2014; Kaminski et al., 2019). Based on these operationalizations, the study

Chapter 1: Introduction

examined whether situation construal predicts performance in different variations of SJT items (situation descriptions, response options, full items), as prior research suggests while both situation descriptions and response options may contain psychologically meaningful information, their individual contribution to situation perception remains unclear. In addition, the study investigated whether this relationship depends on the presence of trait-relevant cues, as suggested by Trait Activation Theory (Tett et al., 2021).

Summary

This chapter has outlined the theoretical foundations and process assumptions underlying SJT performance, with a focus on how test-takers psychologically construe situations. Drawing from person–situation interaction theory and three cognitive process models, a working model was developed that integrates construal components such as trait-relevant cues, ATIC, DIAMONDS, situational strength, and social desirability into three phases of SJT responding. Prior research suggests that while SJTs were originally conceptualized as contextualized simulations, response behavior may be driven just as much or more by other factors, raising questions about what makes SJTs truly “situational.” Despite a wealth of theoretical propositions, empirical clarity remains limited. This dissertation addresses that gap by testing which parts of SJTs activate situation construal processes, which parts are most relevant for construct- and criterion-related validity. The following chapters contribute empirical data to examine the applicability of the working model and inform the ongoing debate about whether or not SJTs primarily measure context-sensitive psychological processes.

References

- Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*(6), 506. <https://doi.org/https://doi.org/10.1037/h0037130>
- Brown, M., & Martin-Raugh, M. (2024). Exploring the Role of Social Desirability in Situational Judgment Tests. <https://doi.org/https://doi.org/10.31234/osf.io/wxqt3>
- Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational Judgement Tests: Constructs assessed and a Meta-Analysis of their criterion-related validities [Review]. *Personnel Psychology*, *63*(1), 83-117. <https://doi.org/10.1111/j.1744-6570.2009.01163.x>
- File, Q. W. (1945). The measurement of supervisory quality in industry. *Journal of Applied Psychology*, *29*(323-337). <https://doi.org/https://doi.org/10.1037/h0057397>
- Freudenstein, J.-P., Remmert, N., Reznik, N., & Krumm, S. (2020). Situational Judgment Test for Teamwork (SJT-TW) [Situational Judgement Test für Teamarbeit (SJT-TA, Gatzka & Volmer, 2017)]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. <https://doi.org/10.6102/zis285>.
- 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. (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>

Chapter 1: Introduction

- Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology*, 60(5), 773. <https://doi.org/10.1037/0022-3514.60.5.773>
- Golubovich, J. (2014). *The impact of situational strength on the validity of situational judgment items*. (Doctoral Dissertation, Michigan State University) ProQuest Information & Learning. <http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2014-99180-083&site=ehost-live>
- Grand, J. A. (2020). A general response process theory for situational judgment tests. *Journal of Applied Psychology*, 105(8), 819. <https://doi.org/10.1037/apl0000468>
- 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 [Editorial Material]. *Industrial and Organizational Psychology-Perspectives on Science and Practice*, 9(1), 23-28. <https://doi.org/10.1017/iop.2015.110>
- Ingold, P. V., Kleinmann, M., König, C. J., Melchers, K. G., & Van Iddekinge, C. H. (2015). Why do situational interviews predict job performance? The role of interviewees' ability to identify criteria. *Journal of Business and Psychology*, 30(2), 387-398. <https://doi.org/10.1007/s10869-014-9368-3>
- Jackson, D. J., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2017). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology*, 90(1), 1-27. <https://doi.org/10.1111/joop.12151>
- Jansen, A., Melchers, K., König, C., Kleinmann, M., Brändli, M., Fraefel, L., & Lievens, F. (2010). Candidates who correctly identify situational demands show better performance. 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA,

Chapter 1: Introduction

- Jansen, A., Melchers, K. G., Kleinmann, M., Lievens, F., 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>
- 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>
- Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology, 78*(6), 988. <https://doi.org/10.1037/0021-9010.78.6.988>
- Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review, 1*(2), 128-146. <https://doi.org/10.1177/2041386610387000>
- König, C., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. C. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment, 15*(3), 283-292. <https://doi.org/10.1111/j.1468-2389.2007.00388.x>
- 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), 399-416. <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. <https://doi.org/10.1037/h0086826>

Chapter 1: Introduction

- Leeds, J. P. (2012). The Theory of Cognitive Acuity: Extending Psychophysics to the Measurement of Situational Judgment. *Journal of Neuroscience Psychology and Economics*, 5(3), 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(6), 2215-2225. <https://doi.org/10.3758/s13428-017-0988-1>
- Lewin, K. (1936). *Principles of Topological Psychology*. McGraw-Hill.
- Lievens, F., Harris, M. M., Van Keer, E., & Bisqueret, C. (2003). Predicting cross-cultural training performance: The validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. *Journal of Applied Psychology*, 88(3), 476-489. <https://doi.org/10.1037/0021-9010.88.3.476>
- Lievens, F., & Sackett, P. R. (2012). The Validity of Interpersonal Skills Assessment Via Situational Judgment Tests for Predicting Academic Success and Job Performance [Article]. *Journal of Applied Psychology*, 97(2), 460-468. <https://doi.org/10.1037/a0025741>
- Marshall, M. A., & Brown, J. D. (2006). Trait aggressiveness and situational provocation: A test of the traits as situational sensitivities (TASS) model. *Personality and Social Psychology Bulletin*, 32(8), 1100-1113. <https://doi.org/10.1177/0146167206288488>
- Martin-Raugh, M. P., & Kell, H. J. (2021). A process model of situational judgment test responding. *Human Resource Management Review*, 31(2), 100731. <https://doi.org/10.1016/j.hrmr.2019.100731>
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis [Review]. *Personnel psychology*, 60(1), 63-91. <https://doi.org/10.1111/j.1744-6570.2007.00065.x>

Chapter 1: Introduction

- 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., & 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., & Hupp, K. (2017). *Wie bedeutsam ist situative Urteilsfähigkeit für die Leistung in SJTs?* (Conference Presentation). DGPS Fachgruppentagung AOW, Dresden, Germany.
- Melchers, K. G., Klehe, U.-C., Richter, G. M., Kleinmann, M., König, C. J., & Lievens, F. (2009). "I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance*, 22(4), 355-374. <https://doi.org/10.1080/08959280903120295>
- 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. New York, United States. <https://doi.org/https://doi.org/10.4324/9780203763643>

Chapter 1: Introduction

- Mischel, W. (1977). On the future of personality measurement. *American Psychologist*, 32(4), 246. <https://doi.org/https://doi.org/10.1037/0003-066X.32.4.246>
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2), 246. <https://doi.org/10.1037/0033-295X.102.2.246>
- 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., 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., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology*. (pp. 241-260). Davies-Black Publishing. <http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1997-36506-008&site=ehost-live>
- Motowidlo, S. J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology*, 66(4), 337-344. <https://doi.org/10.1111/j.2044-8325.1993.tb00543.x>
- Ostrom, J. K., Melchers, K. G., Ingold, P. V., & Kleinmann, M. (2016). Why do situational interviews predict performance? Is it saying how you would behave or knowing how you should behave? *Journal of Business and Psychology*, 31, 279-291. <https://doi.org/10.1007/s10869-015-9410-0>
- 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.

Chapter 1: Introduction

83-105). Lawrence Erlbaum Associates Publishers.

<http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2006-00547-005&site=ehost-live>

- 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. <https://doi.org/10.1037/a0037250>
- Rauthmann, J. F., & Sherman, R. A. (2015). Ultra-brief measures for the situational eight DIAMONDS domains. *European Journal of Psychological Assessment, 32*(2). <https://doi.org/https://doi.org/10.1027/1015-5759/a000245>
- 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>
- 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, 87*, 103963. <https://doi.org/10.1016/j.jrp.2020.103963>
- 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., Mussel, P., Lievens, F., König, C. J., Freudenstein, J.-P., & Krumm, S. (2019). The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology, 105*(8), 800–818. <https://doi.org/10.1037/apl0000457>

Chapter 1: Introduction

- Tett, R. P., & 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>
- Tett, R. P., Toich, M. J., & Ozkum, S. B. (2021). Trait activation theory: A review of the literature and applications to five lines of personality dynamics research. *Annual Review of Organizational Psychology and Organizational Behavior*, 8, 199-233. <https://doi.org/10.1146/annurev-orgpsych-012420-062228>
- Wang, D., Oostrom, J. K., & Schollaert, E. (2023). The importance of situation evaluation and the ability to identify criteria in a construct-driven situational judgment test. *Personality and Individual Differences*, 208, 1-10. <https://doi.org/https://doi.org/10.1016/j.paid.2023.112182>
- Webster, E. S., Paton, L. W., Crampton, P. E., & Tiffin, P. A. (2020). Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education*, 54(10), 888-902. <https://doi.org/10.1111/medu.14201>
- 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., & Jones, C. (1999). Further studies of situational tests [Article]. *Personnel psychology*, 52(3), 679-700. <https://doi.org/10.1111/j.1744-6570.1999.tb00176.x>
- Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, 52(5), 372. <https://doi.org/https://doi.org/10.1037/h0026244>

Chapter 1: Introduction

Wiggins, J. S. (1966). Social desirability estimation and "faking good" well. *Educational and Psychological measurement*, 26(2), 329-341.
<https://doi.org/10.1177/001316446602600206>

Wolcott, M. D., Lobczowski, N. G., Zeeman, J. M., & McLaughlin, J. E. (2021). Does the ability to identify the construct on an empathy situational judgment test relate to performance? Exploring a new concept in assessment. *Currents in Pharmacy Teaching and Learning*, 13(11), 1451-1456. <https://doi.org/10.1016/j.cptl.2021.09.003>

Chapter 2






ATIC in SJTs

This article is published under Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License

(<https://creativecommons.org/licenses/by-nc-nd/4.0/>) as:

Reznik, N., Krumm, S., Freudenstein, J. P., Heimann, A. L., Ingold, P., Schäpers, P., & Kleinmann, M. (2024). Does understanding what a test measures make a difference? On the relevance of the ability to identify criteria for situational judgment test performance. *International Journal of Selection and Assessment*, 32(2), 210-224. <https://doi.org/10.1111/ijsa.12458>

Does understanding what a test measures make a difference? On the relevance of the ability to identify criteria for situational judgment test performance

Nomi Reznik¹  | Stefan Krumm¹  | Jan-Philipp Freudenstein²  |
Anna L. Heimann³ | Pia Ingold⁴  | Philipp Schäpers⁵  | Martin Kleinmann³

¹Department of Education and Psychology,
Freie Universität Berlin, Berlin, Germany

²Group R&D, Hogrefe Verlagsgruppe GmbH,
Göttingen, Germany

³Department of Psychology, University of
Zurich, Zurich, Switzerland

⁴Department of Psychology, University of
Copenhagen, Copenhagen, Denmark

⁵Department of Psychology and Sports,
Westfälische Wilhelms-Universität Münster,
Münster, Germany

Correspondence

Nomi Reznik, Department of Education and
Psychology, Division Psychological
Assessment, Differential and Personality
Psychology, Freie Universität Berlin,
Habelschwerdter Allee 45, 14195 Berlin,
Germany.

Email: nomi.reznik@fu-berlin.de

Funding information

Deutsche Forschungsgemeinschaft; German
Research Foundation (DFG),
Grant/Award Number: KR 3457/2-2

Abstract

Situational judgment tests (SJTs) are low-fidelity simulations that are often used in personnel selection. Previous research has provided evidence that the ability to identify criteria (ATIC)—individuals' capability to detect underlying constructs in nontransparent personnel selection procedures—is relevant in simulations in personnel selection, such as assessment centers and situational interviews. Building on recent theorizing about response processes in SJTs as well as on previous empirical results, we posit that ATIC predicts SJT performance. We tested this hypothesis across two preregistered studies. In Study 1, a between-subjects planned-missingness design ($N = 391$ panelists) was employed and 55 selected items from five different SJTs were administered. Mixed-effects-modeling revealed a small effect for ATIC in predicting SJT responses. Results were replicated in Study 2 ($N = 491$ panelists), in which a complete teamwork SJT was administered with a high- or a low-stakes instruction and showed either no or a small correlation with ATIC, respectively. We compare these findings with other studies, discuss implications for our understanding of response processes in SJTs, and derive avenues for future research.

KEYWORDS

ability to identify criteria, planned missingness, situational judgment test

Practitioner points

- Not much is known about the relevance of ATIC for situational judgment tests (SJTs).
- Two studies revealed a small or no effect for ATIC in predicting SJT responses.
- ATIC variance might be explained more by constructs that items tap into than by individuals.

1 | INTRODUCTION

Situational judgment tests (SJTs) are personnel selection tools that have surged in popularity in the past three decades (Motowidlo et al., 1990). SJT items usually consist of a short text describing an—oftentimes work-related—situation, several response options, and an instruction on how to answer the items (Weekley et al., 2015). The recent popularity of SJTs is not surprising considering that they are relatively cost-effective and easy to administer, while offering substantial predictive validity for job performance (Cabrera & Nguyen, 2001; Christian et al., 2010; McDaniel et al., 2007; McDaniel et al., 2001) and positive applicant reactions (Chan & Schmitt, 1997).

Conceptualized as (low-fidelity) simulations, some scholars assume that responding to an SJT might follow similar principles that apply to other simulations in personnel selection, such as assessment center exercises or situational interviews (Motowidlo et al., 1990; Weekley et al., 2015). Specifically, individuals need to understand the simulated situation at hand and decide how they would respond (e.g., Grand, 2020; Rockstuhl et al., 2015). Importantly, the decision on how to best respond may also be guided by an individual's assumptions about the criteria they will be evaluated on. Indeed, the ability to identify criteria (ATIC; Kleinmann, 1993), defined as an individual's capability to see through nontransparent selection procedures and identify the psychological construct that is being assessed (e.g., Kleinmann et al., 2011), was found to be of substantial relevance in assessment centers and situational interviews (Ingold et al., 2015; Jansen et al., 2013; König et al., 2007). However, while the relevance of ATIC is fairly well established for the aforementioned simulations, much less is known about the relevance of ATIC in SJTs (for an exception, see Wang et al. [2023]). In the current paper, we build on prior theorizing as well as on extant empirical evidence (for details, see below) and examine whether ATIC predicts SJT performance across two studies with two different operationalizations of ATIC in SJTs.

In doing so, we make several contributions. First, we shed more light on the processes underlying SJT responses and thereby add to the ongoing debate on SJT functioning and their construct validity (e.g., Lievens & Motowidlo, 2016). Second, we also contribute to a deeper understanding of SJTs' criterion-related validity. Note that ATIC has been identified as a contributor to the criterion-related validity of assessment centers and situational interviews (e.g., Ingold et al., 2015). Third, we transfer research that has proven insightful in the realm of assessment centers and situational interviews to SJTs as another simulation method. This will help identify common principles in simulations and ultimately contribute to a more holistic view on personnel selection methods (Lievens & Sackett, 2017).

2 | THEORETICAL BACKGROUND

The traditional view of how responses to SJTs are formed is that test-takers visualize the situation described in the item, imagine themselves acting in the situation, and choose a response option

that aligns with their judgment on how to act in the given situation (Motowidlo et al., 1990; Weekley et al., 2015). The processes of interpreting the situation in SJTs has been addressed in several studies. For instance, Rockstuhl et al. (2015) administered an SJT on intercultural interactions in a constructed response format. In addition to test-takers' responses on how they would act in a given situation (response judgment), they were also asked to judge the situations per se, which Rockstuhl et al. (2015) referred to as *situational judgment*. Rockstuhl et al. (2015) revealed, across two studies, that the quality of test-takers' situational judgment was significantly correlated with the quality of their response judgment ($r = .48$ and $r = .49$; Studies 2 and 4, respectively). Contrary evidence, however, was presented by Krumm et al. (2015) as well as Schäpers et al. (2019, 2020). These authors presented findings suggesting that situation descriptions in SJTs had little relevance for SJT performance. These insights were further differentiated by Freudenstein et al. (2020), who addressed *situation construal* in SJTs, defined as an individual's subjective perception of the situation. In a series of studies, they asked participants to report their situation construal in terms of the DIAMONDS framework (Rauthmann et al., 2014). In line with Rockstuhl et al. (2015) these authors found that test-takers' situation construal was relevant for their responses in an SJT item—but that relevant situation construal was mostly driven by SJTs' response options.

The situated reasoning and judgment framework (SiRJ; Grand, 2020) offers further explanation of SJT functioning. Within the SiRJ framework, the cognitive processes involved in SJT responding are broken down into *conditional reasoning*, *similarity judgments*, and *preference accumulations*. Conditional reasoning is said to occur during the process of reading, understanding, and interpreting the SJT item. So, conditional reasoning refers to test-takers' perception of the situation and their decision about the situational demands that should be acted upon. As such, this process is thought to create a frame of reference for subsequent similarity judgments. Similarity judgments refer to comparisons between test-takers' self-generated behavior and behavior described in each of the response options. Finally, preference accumulations denote the process of deciding which response option best matches test-takers' preferred choice of action. Grand (2020) applied a computational approach and found that the response processes of simulated test-takers converged with the assumptions made in the SiRJ model. Thus, several studies converge in that *understanding the situation at hand*—referred to as either situational judgment (Rockstuhl et al., 2015) or situation construal (Freudenstein et al., 2020) or conditional reasoning (Grand, 2020)—is an important determinant of SJT functioning. We hereinafter use the term 'situational judgment' to refer to the process of understanding situations in SJT items in general. When referring to specific studies or models, we use the terminology adopted by the respective authors.

Interestingly, research on other simulations (i.e., assessment centers and situational interviews) took a somewhat different but related route to examining the situational judgment that may be relevant. Specifically, in the realm of assessment centers and

situational interviews, ATIC has been identified as a relevant driver of performance (Ingold et al., 2015; Klehe et al., 2012; Kleinmann et al., 2011). ATIC refers to test-takers' *ability* to understand cues in personnel selection procedures (Kleinmann, 1993). Specifically, ATIC addresses test-takers' ability to identify what a given selection procedure demands of them, which then helps them to align their responses to these demands. As such, ATIC builds on the premise that participants try to present themselves positively in personnel selection procedures (Melchers et al., 2009). To achieve this, participants make assumptions about the criteria that will be used to evaluate their performance. ATIC thus refers to the extent to which these assumptions are correct (i.e., converge with the actually relevant criteria).

The empirical relevance of ATIC in assessment center exercises has been well established (for an overview see Kleinmann and Ingold [2019], Kleinmann et al. [2011], König et al. [2007]). In multiple studies, participants first completed an assessment center exercise. Subsequently, they were asked what construct (i.e., dimension) they believe was assessed in the exercise. Their answers were then rated for correctness by trained coders. In such studies, ATIC has been shown to significantly predict assessment center performance, with correlations ranging from $r = .23$ to $r = .49$ (Jansen et al., 2013). Similar results were reported for situational interviews (Ingold et al., 2015; Oostrom et al., 2016). In these studies, a situational interview was conducted first. Subsequently, interviewees were shown the exact same interview questions and were then asked to name the targeted constructs. Their ATIC performance (i.e., the correspondence between the interviewees' assumption and the actual target construct, as determined by trained coders) showed substantial correlations with the actual interview performance as well as with performance in a simulated work setting (Oostrom et al., 2016) and with supervisory ratings of job performance over and above interview performance (Ingold et al., 2015).

Importantly, the conceptualization of ATIC as the ability to understand and identify relevant task-related information in an assessment situation (Kleinmann, 1993) shares similarities with—but is not identical to—situational judgment in SJT items. Note that Rockstuhl et al. (2015, p. 465) defined situational judgment “as individuals' sense-making of a situation, which enables them to comprehend, explain, attribute, extrapolate, and predict situations.” Similarly, conditional reasoning in the SiRJ model is defined as perceiving a situation and deciding which of its demands should be acted upon. Thus, ATIC is similar to the aforementioned concepts in that it also refers to the identification and, respectively, understanding of situational demands. However, ATIC adopts a different reference point. While the situational judgment in SJTs describes how an individual perceives and categorizes a given situation, ATIC describes whether an individual correctly recognizes on which criteria they are evaluated in different situations (see also Wang et al., 2023). That is, ATIC is conceptualized as an individual ability (Kleinmann et al., 2011) whereas situational construal in SJTs and conditional reasoning have so far not been discussed as stable interindividual differences across several situations. It may thus be inferred that

ATIC might be a predictor of individuals' situation construal in an assessment situation (along with other person characteristics and determinants of the situation), which also aligns well with assumptions made by situation construal models (e.g., Funder, 2016).

The notable differences between ATIC and situational judgment notwithstanding, SJTs are similar to assessment center exercises and situational interviews in many ways (Jansen et al., 2013; Lievens, 2006), thus suggesting that ATIC is also relevant in SJTs. First, all of these methods are considered simulations (e.g., Motowidlo et al., 1990). As such, they ask participants to envision themselves in and respond to simulated situations. If participants aim at presenting themselves as positively as possible, that is, showing maximum performance, a necessary requirement for them is to identify the relevant criteria and act accordingly. Second, all of these selection methods feature nontransparent situations, albeit to a different degree (see below). Participants are usually not informed which construct is being assessed. However, participants will be able to present themselves more favorably in a selection procedure, if they can correctly anticipate the construct being assessed—which is referred to as ATIC—and respond accordingly (Kleinmann et al., 2011). Third and relatedly, all of these methods build on the concept of behavioral consistency. That is, behavior in a simulated workplace situation is similar to and predictive for behavior in a real-life workplace situation (Thornton & Cleveland, 1990). Thus, ATIC will not only be a means to achieve good scores in an assessment but also in real work situations outside the assessment context (Jansen et al., 2013). These similarities suggest that ATIC may be a relevant—but, of course, not the only—driver of performance, not only in assessment centers and situational interviews (Ingold et al., 2015; Kleinmann et al., 2011), but also in SJTs.

Only a few studies have so far provided empirical evidence on the relationship between ATIC and SJT performance. Oostrom et al. (2016) administered 24 video SJT items with a behavioral response format (i.e., test-takers were asked to respond directly through a web-cam to video-taped actors) and additionally gauged test-takers' assumption about the measurement intention of the SJT items. The correctness of these assumptions, rated by independent researchers, served as an ATIC score. Average performance ratings for behavioral responses to SJT items correlated around .43 with ATIC scores. In an unpublished study, Melchers and Hupp (2017) applied SJTs in a more common format (i.e., written scenarios and a closed response format). Their study yielded a correlation of $r = .38$ between test-takers' SJT scores (aggregated per test-taker as a single score across multiple SJTs) and ATIC scores. Conversely, Wolcott et al. (2021) administered an empathy SJT to a small sample and found no relationship between ATIC and SJT performance. The most comprehensive research on the relevance of ATIC in SJTs was presented by Wang et al. (2023). Across three studies, these authors found that (a multiple-choice measure of) ATIC was significantly related to performance in a construct-driven SJT (standardized $\beta = .29$ and $\beta = .31$). Paralleling findings from assessment centers and situational interviews, they also revealed that ATIC provided an incremental prediction (above and beyond SJT performance) of an interpersonal performance criterion.

Hence, for the majority of previous studies, ATIC has shown a medium-sized effect on SJT performance (according to established effect size conventions, see Cohen [1992]). We thus propose the following hypothesis (as preregistered; Study 1: https://osf.io/b6e9s/?view_only=fbee70aed8434e169acb03ec1bda736d and Study 2: https://osf.io/2yter/?view_only=fd131aefb6984d97af4a53a7e1c4737e):

Hypothesis 1. ATIC will predict SJT performance. This effect will be positive and of moderate size.

The current preregistered studies will test this hypothesis. In doing so, we exceed previous research on the relevance of ATIC in SJTs in several ways. First, we use prototypical traditional and construct-driven SJTs. Second, we employ a broad set of items from several SJTs. Third, we use the traditional and most common way of assessing ATIC (as opposed to the multiple-choice ATIC measure used by Wang et al. [2023]). Moreover, we analyze data on the SJT item level (Study 1), thereby accounting for the multidimensionality that is typically evident between and within SJTs (Tiffin et al., 2020; Whetzel et al., 2020) as well as on the test level (Study 2), thereby following the more typical approach of other studies on ATIC (Ingold et al., 2015; Klehe et al., 2012; Kleinmann et al., 2011).

3 | STUDY 1

3.1 | Method

3.1.1 | Participants

A total of 450 participants (for sample size recommendations, see Green [1991]) took part in the study. Participants were recruited via the online panel provider prolific.co and incentivized with a payment of £8 per hour. Fifty-nine participants were excluded for careless responding (e.g., giving the same response to all ATIC questions) and/or giving nonsense responses (e.g., entering random letters or numbers as responses to ATIC questions, such as “aaaaaaaa” or “123456”), resulting in 391 participants ($f = 171$, other = 1) being included in the statistical analyses. On average, participants were 27.97 years old ($SD = 8.83$). A proportion of 54.8% held a university degree, 58.8% were employed full-time, and an additional 35.1% were currently studying and holding a part-time job.

3.1.2 | Study design and materials

Initial SJT item selection. As little research had yet been conducted on ATIC in SJTs, we chose to include SJT items from multiple tests to maximize the generalizability of our findings. Hence, we chose SJTs based on their typicality and sought to cover the construct domains of applied social skills and personality (i.e., the construct domains that cover the majority of SJTs; Christian et al., 2010). As a result, our

initial item pool consisted of 78 items from five different SJTs: (1) the personal initiative SJT (Bledow & Frese, 2009), (2) a translated version of the SJT for teamwork (Gatzka & Volmer, 2017; translated by Freudenstein et al. [2020]), (3) the team role test (Mumford et al., 2008), (4) the SJT for employee integrity (Becker, 2005), and (5) the HEXACO SJT (Oostrom et al., 2019). Note that while we chose items from five different SJTs, the herein included items covered 10 different constructs (e.g., items from the HEXACO SJT covered several personality constructs).

Items from these SJTs were inspected by subject matter experts (SMEs; one author of this study as well as three research assistants with at least 1 year of experience in the research field) to ensure that they were suitable as ATIC assessments. SMEs received a briefing and completed several training items before they were randomly assigned to rate the following criteria. Two of the SMEs assessed which construct was assessed by each item and whether an item tapped into one singular construct. In other words, we made sure that only one correct ATIC response existed. Notably, SMEs' judgment concerning the targeted constructs converged with the constructs that were intended by the test authors for all of the items, which speaks to the construct validity of the items.¹ The remaining two SMEs independently checked whether the item contained specific and understandable information about a work-related situation (i.e., item clarity; adapted from Meyer et al. [2014]).

If items were not rated as tapping into one single construct and/or as not being sufficiently clear, they were excluded.² This was the case for 23 items (see Mussel et al. [2018], Tiffin et al. [2020] for similar problems with SJT items), resulting in a final item pool of 55 items from five different SJTs (see Supporting Information: Appendix B).

Study design. We implemented a planned missingness three-form design (Rhemtulla & Hancock, 2016) to reduce participant burden. That is, we divided items across four item sets: an X-set which all participants completed, as well as A-, B-, and C-sets, to which participants were randomly assigned. The X-set comprised two items from each of the five SJTs, which were either chosen based on the highest item-total correlations (if such data was available) or randomly assigned. The remaining 45 items were randomly sorted into A-, B-, and C-sets. Thus, each participant completed a total of 25 items (which is in line with the average length of typical SJTs; we counted an average length of 23 items among an ad-hoc selection of 15 prominent SJTs).

First, participants were instructed to imagine that they were currently applying for their dream job and to imagine that this survey was part of their application process. Participants then completed all SJT items of the X-set and of the randomly assigned A-, B-, or C-set. All SJT items were presented using a behavioral tendency (“would-do”) response instruction (McDaniel et al., 2007). Participants responded to SJT items by rating each individual response option on a 7-point rating scale (1 = *do not agree*, 7 = *fully agree*). After responding to 25 SJT items, we assessed ATIC by again presenting the same SJT items participants had just completed (with no indication on how they had responded to the item). This time,

participants were asked to specify in an open response format and for each SJT item, which construct they thought had been targeted and to provide behaviors they associated with the construct (following the typical routine of assessing ATIC; see Kleinmann [1993], Kleinmann et al. [2011]). An example item for assessing in ATIC in SJTs is presented in Supporting Information: Appendix A.

Scoring. To score SJT item responses, which were made on a rating scale from 1 to 7, we employed a distance scoring method. We deducted seven points from participants' ratings if the response option was listed (by the test authors) as correct or as indicative of a high standing on the targeted trait. This means that participants who rated such a response as 7 (= "fully agree") received a score of 0, indicating no deviation from the "ideal" response. Conversely, participants who rated such a response as 1 (= "do not agree") received a score of -6, indicating a maximum deviation from the ideal response.

We deducted one point if the response option was listed as incorrect or indicative of a low standing on the targeted trait. This means that participants who rated an incorrect a response as 1 (= "do not agree") received a score of 0, indicating no deviation from the ideal response. Conversely, participants who rated such a response as 7 (= "fully agree") received a score of +6, again indicating a maximum deviation from the ideal response.

We deducted four points (i.e., the midpoint of possible scores on a scale from 1 to 7) if the response option was listed as neither correct nor incorrect or neither indicative of a high nor low standing on the targeted trait. In doing so, we followed the recommendation of several authors of the herein included SJTs (e.g., Becker, 2005; Bledow & Frese, 2009; Gatzka & Volmer, 2017). This means that participants who indicated that a response was neither correct nor incorrect (by choosing the midpoint of our scale of 4) received a score of 0, thus indicating no deviation from the ideal response.

Overall, the resulting values ranged from -6 to +6 and were then converted into absolute values, thus ranging from 0 (*best score*) to 6 (*worst score*).³ Thus, the final scores reflected the absolute distance from the ideal response (Whetzel et al., 2020; Wolcott et al., 2019, for a similar procedure, see Oostrom et al. [2012]).

As preregistered, ATIC responses were scored by two SMEs. They independently rated whether participants' responses aligned with the measurement intention of the SJTs' authors on a 4-point scale ranging from 0 ("does not fit the construct at all") to 3 ("fits the construct perfectly"), which follows the established way of scoring ATIC (e.g., Ingold et al., 2015).⁴

Interrater reliabilities for all ATIC response ratings were assessed per item. Initially, 17 out of 55 items showed intraclass correlation coefficients (ICCs) below .50 (i.e., below moderate agreement, see LeBreton and Senter [2008]). In these instances, issues regarding different understandings of the specific answers were discussed among raters, thus following the current approach to ATIC ratings (see Ingold et al., 2015; Kleinmann et al., 2011). After this reassessment, ICCs for all items ranged from .53 to .97 with a mean ICC of .76.

3.1.3 | Analytic strategy

To predict SJT item responses, we applied linear mixed-effects regression with crossed random intercepts on Level 2 (Baayen et al., 2008), as item responses were clustered both within persons and within items. Specifically, we fitted a mixed-effects model with random intercepts for the SJT item, the individual, and their interaction, respectively, and random slopes for ATIC per individual and per SJT item. We used R-packages lme4 (Bates et al., 2015, Version 1.1.28.) and lavaan (Rosseel, 2012, Version 0.6.10.) in RStudio (Version 4.1.2; R Core Team, 2021). Missing data, as intended by our study design, was estimated using full-informed maximum likelihood (Hox et al., 2010).

We emphasize that we sampled a plethora of SJT items across SJTs and then carefully selected out items. The SJT items that survived this selection were suitable for an ATIC assessment. As a result, the remaining items did not represent complete SJTs and thus did not warrant being aggregated to test scores. However, the herein adopted approach in analyzing the relationship between ATIC and SJT response behavior enabled us to account for variance that was due to participants' individual differences and due to differences across SJT items.

3.2 | Results

3.2.1 | Preliminary analyses

Across all SJT items and all individuals, mean ATIC performance was 0.62 (on a scale ranging from 0 to 3, $SD = 0.30$). Note that this is rather low for ATIC performances gauged in the context of assessment centers or situational interviews (see values around 1.50 at Ingold et al. [2015]) but similar to ATIC performances reported for SJTs (Melchers & Hupp, 2017; Oostrom et al., 2016). We also observed that ATIC scores differed across SJT items. That is, we found high ATIC mean scores (of up to 2.50) for some items measuring honesty-humility (HEXACO-SJT; Oostrom et al., 2019) and employee integrity (Becker, 2005). Conversely, items measuring team roles (team role test; Mumford et al., 2008) and personal initiative (Bledow & Frese, 2009) showed low ATIC scores (scores of 0.10 or even 0.03).

When inspecting bivariate correlations, we found significant correlations between SJT item performance and ATIC performance only for some of the items (see Table 1 for items of the X-set, see Supporting Information: Appendix C for all remaining items). Notably, all correlations were generally on the lower side, with an average of $r = -.01$. Note that due to the distance scoring of SJT items, negative correlation values reflected a positive relation between SJT item performance and ATIC responses. However, we also revealed positive correlation coefficients for some SJT items with ATIC responses. We also observed significant correlations between SJT items with ATIC scores from different SJT items (see Table 1, correlations above and below the diagonal). Given the seemingly random pattern and the large amount of correlations (cf. the risk of alpha inflation), we refrain from interpreting these results in more detail.

Chapter 2: ATIC in SJTs

TABLE 1 Means, standard deviations, and bivariate correlations between ATIC scores and SJT item scores.

ATIC scores	SJT itemscores											
	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10
1. Employee integrity SJT (Item 2)	0.27	0.51	-.08	-.04	.07	.03	-.01	.03	-.00	.01	.05	-.06
2. Employee integrity SJT (Item 5)	0.41	0.61	.08	-.17**	.01	.06	-.04	-.05	-.01	.09	-.09	-.03
3. HEXACO SJT (Item 7, H)	2.18	1.06	-.08	-.10	.00	-.03	-.09	-.13*	-.01	-.06	-.01	-.11*
4. HEXACO SJT (Item 13, H)	2.13	1.08	-.02	-.16**	.01	-.07	-.06	-.08	-.03	.01	.00	-.13*
5. Personal initiative SJT (Item 6)	0.12	0.43	-.03	.04	.01	-.08	.04	-.05	-.01	-.06	-.02	-.01
6. Personal initiative SJT (Item 12)	0.03	0.21	-.02	-.05	.02	.07	.00	-.08	-.04	.10	.05	.02
7. Team role SJT (Item 6)	0.68	0.81	-.06	-.12*	-.06	.02	-.01	-.01	-.07	.08	-.12*	-.17**
8. Team role SJT (Item 8)	0.45	0.58	-.07	-.02	-.04	.09	.03	.05	-.02	.00	-.08	-.14**
9. Teamwork SJT (Item 2)	0.69	0.92	-.08	-.07	-.04	.02	-.14*	-.05	.03	.11*	-.13*	-.14**
10. Teamwork SJT (Item 5)	0.42	0.80	-.02	-.05	-.05	.02	.03	-.00	-.06	.04	-.08	-.10

Note: Item numbers are presented to reflect the number assigned in the original SJTs: teamwork SJT (Gatzka & Volmer [2017] translated by Freudenstein et al. [2020]), personal initiative SJT (Bledow & Frese, 2009), team role SJT (Mumford et al., 2008), employee integrity SJT (Becker, 2005), HEXACO SJT (Oostrom et al., 2019). Negative correlation coefficients reflect a positive relation between SJT item performance and ATIC responses.

Abbreviations: ATIC, ability to identify criteria; H, honesty/humility; SJT, situational judgment test.

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

Since mean ATIC performance varied between SJT items measuring different constructs, we also inspected bivariate correlations among SJT items aggregated per each construct. We did this for subsets from the A-, B-, and C-sets only since these sets contained more items than the X-set that could be aggregated. Descriptively, the strongest correlations between SJT and ATIC performance were observed for teamwork items ($r = -.25$, $r = -.33$, and $r = -.14$) and employee integrity items ($r = -.26$, $r = -.12$, and $r = -.21$, for A-, B-, and C-sets, respectively). Correlations among aggregated scores for all other constructs showed correlations at or below $|.10|$ and were on average around zero.

3.2.2 | Hypothesis test

To test our hypothesis—whether ATIC predicted SJT performance with a moderate effect size—we applied linear mixed-effects regression with crossed random intercepts on Level 2 (Baayen et al., 2008) and used SJT item responses as dependent variables. Model comparison revealed that the random-slope model fitted the data better than the intercept-only model. However, the random slopes did not explain substantial variance (see Table 2). The fixed-effects estimate for ATIC on the individual level, that is, the overall effect of ATIC on SJT responses across all items, was significant (estimate = -0.15 , confidence interval [CI] = -0.24 to -0.06 , $p = .001$),⁵ meaning that higher ATIC scores by one point (on a scale ranging from 0 to 3) lead to SJT responses that were 0.15 closer to the correct solution. With the standard deviation for the SJTs being 0.78, we consider this a small effect. The fixed-effects estimate for ATIC on the item-level (i.e., the situation-specific deviation from the individual effect) was not significant (estimate = -0.02 , CI = -0.05 to 0.00 , $p = .06$). Thus, our hypothesis was partially supported, in that

TABLE 2 Results of mixed-effects model for ATIC predicting SJT responses.

Predictors	Estimates	CI	<i>p</i> -Value
Fixed effects			
(Intercept)	2.15	2.03-2.26	<.001
ATIC (item)	-0.02	-0.05 to 0.00	.06
ATIC (individual)	-0.15*	-0.23 to -0.05	.001
Random Effects			
σ^2	2.14		
τ_{00} individual	0.03		
τ_{00} SJT Item	0.13		
τ_{11} SJT Item (ATIC)	0.00		
ρ_{01} SJT Item	-.04		
ICC	0.10		
Marginal R ² / Conditional R ²	0.001 / 0.085		

Note: $N = 391$.

Abbreviations: ATIC, ability to identify criteria; CI, confidence interval; ICC, intraclass correlation coefficient; SJT, situational judgment test.

ATIC performance significantly predicted SJT performance, but with a small instead of the hypothesized moderate effect.

3.2.3 | Ancillary analyses

Since we observed that ATIC scores varied substantially across SJT items, we conducted an ancillary analysis to identify what accounted for this variability. Using a G-theory-based approach

(e.g., Woehr et al., 2012), we revealed that variance in ATIC was largely due to the constructs (e.g., integrity, teamwork, honesty/humility) measured by the SJT items (25.3% of variance), whereas only 4.4% of variance in ATIC was due to individuals, and 10.3% of variance was due to the interaction of individual and construct. When we reran this analysis with the SJT test as a variance component instead of SJT constructs, the amount of explained variance by SJT tests was only 10.0%.

3.3 | Discussion

Study 1 examined the relationship between ATIC and SJT performance on the item level. Across 55 items that tapped into 10 different constructs and came from five different SJTs, we found partial support for our hypothesis. We revealed that the herein included items yielded only a small relationship between ATIC and SJT performance. Mixed regression analysis attested that this was true on the individual level and on the item level. That is, individuals scoring higher on ATIC (across all SJT items) tended to show only slightly better SJT item responses (and vice versa). Moreover, items in which a better ATIC score was achieved did not yield better SJT item responses (and vice versa). The latter is particularly surprising considering that ATIC varied substantially across items—and much less across individuals—but still did not significantly explain SJT performance on the item level. So, a preliminary conclusion from Study 1 is that—contrary to findings in the realm of assessment centers and situational interviews (Ingold et al., 2015; Jansen et al., 2013; Kleinmann et al., 2011; König et al., 2007)—ATIC may be of little relevance in the herein applied set of SJT items.

However, several design choices may have influenced results of Study 1 and call for further research. First, we examined the relationship between ATIC and SJT performance on the item level (in contrast to Melchers & Hupp [2017], Wang et al. [2023]). Administering a variety of SJT items across several constructs was beneficial for the generalizability of our results across items and constructs. On the other hand, these results may not generalize to typical SJTs in which a series of similar items is presented consecutively. As a result of the design choice to include items from several constructs, a behavioral tendency response instruction was the only option that worked for all SJT items (as knowledge response instructions were not appropriate for SJTs tapping into the personality domain). However, a behavioral tendency response instruction may pose a disadvantage for ATIC to predict SJT performance since participants were not specifically prompted to find responses that are most effective in real settings. This may be especially true since Study 1 did not include a performance incentive.⁶

To address these issues, we conducted Study 2. In line with previous studies on ATIC in SJTs, Study 2 applied (i) an entire SJT (on teamwork, which is a construct i.e., frequently targeted by SJTs; Christian et al., 2010) with (ii) a knowledge instruction. To illuminate whether the effects of ATIC on SJT performance may be prone to (high- vs. low-incentive) framing effects, we furthermore

implemented (iii) a between-subjects condition in which we manipulated the incentives that were available for good SJT performances.

4 | STUDY 2

4.1 | Method

4.1.1 | Participants

The data collection in the study (as preregistered; https://osf.io/2yter/?view_only=fd131aefb6984d97af4a53a7e1c4737e) followed the recommendations of Schönbrodt and Perugini (2013) to collect 250 participants for stable correlations. A total of $N = 510$ participants from the online panel prolific.co completed the online questionnaire. Participants were incentivized with £4.45 for participating in the study, which took approximately 37 min to complete. Nineteen participants were excluded since they either failed to correctly answer two careless responding check items (Meade & Craig, 2012) or indicated that their data should not be used for further analyses, resulting in 491 participants ($f = 206$, other/diverse = 4; no gender indicated = 74) being included in the statistical analyses. The final sample had a mean age of 29.63 years ($SD = 8.92$). For analyses on the test level, we additionally excluded 57 and 36 participants who gave nonsense responses to one or more ATIC questions (e.g., random letters or numbers) for the high- and low-incentive sample, respectively.⁷

4.1.2 | Study design and materials

Situational judgment test. To expand on our findings from Study 1, we administered a full SJT. We chose the translated version of the Teamwork SJT (Gatzka & Volmer (2017); translated by Freudenstein et al. [2020]), as its items showed the highest bivariate correlations between ATIC and item performance in Study 1. The Teamwork SJT consists of 12 items with a knowledge response instruction (“What should your team do and not do in such a situation?”) as well as a pick-the-best-and-the-worst response format. Reliability estimates were low in both conditions (omegas = 0.48 and 0.51 for low- and high-incentive conditions, respectively), which is typical for many SJTs (e.g., Kasten & Freund, 2016) and in line with other studies using the teamwork SJT (Freudenstein et al., 2023).

Test motivation. To rule out that SJT and ATIC responses were lowly correlated because of the insufficient motivation of participants, we administered three items tapping into the test motivation factor of the Test Attitude Survey (Arvey et al., 1990, e.g., “Doing well on this test was important to me”). Participants responded on a scale from 1 to 5. Reliabilities were acceptable (omegas = 0.73 and 0.71 in the high- and low-incentive conditions).

Study design. To further expand the findings from Study 1, we employed a between-subjects design. That is, we randomly assigned participants to either a condition in which a similar instruction was

Chapter 2: ATIC in SJTs

given as in Study 1 (i.e., to imagine that this test was part of a selection procedure) or a condition in which they were additionally offered a bonus payment of £25 if their performance in the SJT was among the best 10% (for a similar procedure see Oostrom et al. [2016]). We hereinafter refer to these conditions as low- versus high-incentive. Apart from the initial instruction, all other aspects were identical in both conditions.

We first presented the 12 items of the teamwork SJT with the original knowledge response instruction as well as the original response format. After this part, we gauged ATIC in the same way as done in Study 1 (i.e., participants saw with the same SJT items again and were asked to specify which construct they believed was being assessed by each item). Only when assessing ATIC, we added three items from another SJT (measuring personal initiative; Bledow & Frese, 2009)—to make the ATIC responses potentially less repetitive and keep participants attentive. These items were not scored and thus not included in the analyses. Finally, participants answered three items measuring test motivation (Arvey et al., 1990), and gave their consent to use their data.

4.1.3 | Scoring

We used the original scoring method as described by the test authors to score SJT items: For each item, participants were asked to pick the best and the worst answer from the response options. Correctly identifying both the best and the worst response was scored with one point each, while wrongly identifying the best and worst response was scored with a negative point each. Thus, item scores ranged from -2 to 2 . Item scores were added across the entire Teamwork SJT, resulting in test scores potentially ranging from -24 to 24 (Gatzka & Volmer, 2017).

To score ATIC, we followed the routine described in Study 1. That is, we utilized the same 4-point-coding scheme ($0 = \text{does not fit the construct at all}$ to $3 = \text{fits the construct perfectly}$). Four independent raters coded all ATIC responses of 50 participants. The remaining participants were then split between raters, resulting in two ratings per ATIC response. Interrater reliabilities for all ATIC response ratings were assessed per item. ICCs for all items ranged from $.84$ to $.97$, with a mean ICC of $.93$.

4.2 | Results

4.2.1 | Preliminary analyses

We first examined mean differences in test motivation, SJT scores, and ATIC scores across both study conditions. We found that test motivation scores were significantly higher in the high-incentive ($M = 4.59$; $SD = 0.60$) than in the low-incentive group ($M = 4.31$; $SD = 0.63$), $t(430.41) = 6.08$, $p < .01$, with a medium-sized effect ($d = 0.56$). Likewise, SJT scores were significantly higher in the high-incentive ($M = 10.08$; $SD = 4.14$) than in the low-incentive group

($M = 8.67$; $SD = 4.24$), $t(478.93) = 3.79$, $p < .01$, with a small effect ($d = 0.34$)—thus indicating that our high- versus low-incentive manipulation had worked. ATIC scores, however, did not differ significantly between high- ($M = 0.56$; $SD = 0.46$) and low-incentive groups ($M = 0.54$; $SD = 0.47$), $t(393.71) = 0.20$, $p = .79$.⁸ Note that we followed common approaches in ATIC research and only incentivized performance in the actual assessment but not ATIC performance (Ingold et al., 2015; Jansen et al., 2013; Melchers et al., 2009). Hence, similar ATIC scores in the two study conditions are in line with this design choice. As in Study 1, mean ATIC performance was rather low but similar to ATIC performances reported for SJTs (Melchers & Hupp, 2017; Oostrom et al., 2016).

4.2.2 | Hypothesis test

While the main goal of Study 2 was to examine correlations of ATIC and SJT responses on the test level, we also inspected correlations on the item level to enable a comparison with Study 1. Item-level results are presented in Tables 3 and 4. For the low-incentive conditions, item-level correlations were mostly around zero and ranged from $-.12$ to $.19$. For the high-incentive condition, a similar pattern was observed (range = $-.20$ to $.14$). On the test level, SJT test scores and mean ATIC scores correlated at $r = .20$ ($p < .01$) and $r = -.06$ ($p = .43$) in the low- and high-incentive conditions, respectively (these correlations changed only marginally when we controlled for test motivation). A comparison of both correlation coefficients revealed that ATIC and SJT performance were significantly more strongly related in the low-incentive condition than in the high-incentive condition ($z = 2.53$, $p = .006$). In sum, we again found partial support for our hypothesis. While ATIC performance significantly predicted SJT performance in one of the two conditions, it did so with a small instead of the hypothesized moderate effect.

4.3 | Discussion

Study 2 sought to scrutinize the results from Study 1 while making crucial changes to the study design. Therefore, we administered a full SJT (on teamwork) with 12 items, employed a knowledge response instruction, and added a condition in which we provided a performance incentive. Correlations between ATIC and SJT scores were largely similar to those obtained in Study 1: The average item-level correlations were around zero in Study 1 as well as in both conditions in Study 2. Interestingly, this is in contrast to our assumptions and the logic behind the design changes made from Study 1 to Study 2. Specifically, we presumed that the behavioral tendency instruction and the absence of a performance incentive had posed a disadvantage to ATIC becoming relevant when responding to an SJT under said conditions (i.e., in Study 1). We speculate that identifying the targeted construct is generally rather difficult in SJTs and cannot be improved through higher incentives (see also the similar ATIC levels in both conditions of Study 2). Moreover, the

Chapter 2: ATIC in SJTs

TABLE 3 Means, standard deviations, and bivariate correlations between ATIC scores and SJT item scores in the low-incentive condition.

ATIC scores	SJT item scores													
	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11	12
SJT Item 1	0.31	0.79	.14*	.16*	-.06	.08	.03	.03	.00	-.03	.00	-.09	.06	-.06
SJT Item 2	0.90	1.11	-.03	.05	.01	.15*	.04	.07	.11	-.08	.02	.17*	.15*	.10
SJT Item 3	0.31	0.82	-.03	.03	-.00	.04	.11	.01	.04	-.07	-.00	.08	.13	-.01
SJT Item 4	0.60	0.89	.06	.04	.05	.03	-.01	.10	-.02	-.06	.03	.10	-.03	-.00
SJT Item 5	0.30	0.79	.11	-.01	.06	.13	.02	.08	.07	.12	.08	-.04	.09	-.02
SJT Item 6	0.46	0.93	.19**	.02	.00	.08	.17*	.09	-.09	.04	.02	-.02	.05	-.12
SJT Item 7	0.47	0.91	.10	.14*	-.04	.07	.18**	.10	.01	.06	.07	.03	.00	-.00
SJT Item 8	0.96	1.17	-.06	.02	-.13	.06	.08	-.03	-.01	.09	-.04	-.00	.00	.02
SJT Item 9	0.67	1.07	.11	.12	-.04	.06	.18**	.11	.01	-.06	-.01	.04	.04	-.03
SJT Item 10	0.46	0.90	.01	.09	-.05	.02	.12	.12	-.07	.00	-.05	.05	.02	-.10
SJT Item 11	0.58	0.90	-.01	.01	.11	.10	.03	.02	.02	.04	-.06	.04	-.01	-.11
SJT Item 12	0.59	1.07	.06	.15*	.03	.12	-.05	.09	-.08	.08	-.00	.04	.02	.05

Note: Item numbers are presented to reflect the number assigned in the original teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. [2020]).

Abbreviations: ATIC, ability to identify criteria; SJT, situational judgment test.

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

TABLE 4 Means, standard deviations, and bivariate correlations between ATIC scores and SJT item scores in the high-incentive condition.

ATIC scores	SJT item scores													
	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11	12
SJT Item 1	0.3	0.74	.08	.00	.01	-.02	-.15*	.04	.08	-.01	-.12	.02	.08	-.06
SJT Item 2	0.95	1.24	.03	-.02	.06	-.01	-.05	.08	.11	-.03	-.02	.07	-.00	-.08
SJT Item 3	0.27	0.76	-.01	-.06	.05	-.06	.06	.09	.00	.04	-.01	.05	.07	-.08
SJT Item 4	0.6	0.91	.03	-.10	-.02	-.06	-.15*	.01	.05	-.17**	-.15*	-.08	.03	-.01
SJT Item 5	0.31	0.82	-.05	.10	-.04	.08	-.04	.07	.06	.13*	-.02	.13	-.01	.04
SJT Item 6	0.44	0.95	-.00	-.05	-.14*	-.08	.01	.11	-.02	-.08	-.10	.07	-.05	.01
SJT Item 7	0.43	0.83	.06	.03	.05	-.10	-.15*	-.12	.01	.09	-.20**	.07	.08	-.03
SJT Item 8	0.95	1.2	-.08	.08	.04	-.11	-.04	.01	.03	.06	-.19**	.08	-.04	-.07
SJT Item 9	0.85	1.2	-.12*	-.00	-.04	.02	.01	-.01	-.14*	.11	-.13*	.04	.01	.04
SJT Item 10	0.35	0.82	-.04	.04	.06	.00	-.07	.03	-.03	.07	-.04	.13*	.03	.05
SJT Item 11	0.61	0.87	.08	.13*	-.04	.14*	-.03	-.03	.14*	.09	-.00	.03	.03	-.04
SJT Item 12	0.45	0.95	.11	.02	.07	-.10	-.09	.00	.04	-.11	-.12	-.04	-.05	-.10

Notes. Item numbers are presented to reflect the number assigned in the original teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. [2020]).

Abbreviations: ATIC, ability to identify criteria; SJT, situational judgment test.

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

applicant framing given in Study 1 may have sufficed for participants to respond to SJT items in a should-do- instead of a would-do-manner—thus potentially rendering the distinction between knowledge- and behavioral-tendency instruction arbitrary. Finally, we also conclude that the correlation between ATIC and SJT

performance was not contingent on the number of similar SJT items that are administered (as this number was higher in Study 2 as compared to Study 1). However, we obtained findings that, in part, suggest that test-level analyses of the relationship between ATIC and SJT performance may reduce some noise, which may be inherent in

item-level analyses. This was evident when comparing the average item-level correlation and the test-level correlation between ATIC and SJT performance in the high-incentive condition in Study 2, but not in the low-incentive condition.

Notably, the obtained correlations differed significantly across the two conditions that were realized in Study 2. Contrary to what one might expect, ATIC was not more strongly related to SJT performance in the high-incentive condition. This runs against the notion that, with higher stakes, applicants will be more inclined to look for clues about what is expected from them and adjust their response behavior accordingly (Kleinmann et al., 2011). From the differences in test motivation and SJT scores, we can infer that our manipulation—to provide an additional incentive in one condition—had generally worked. We can only speculate why ATIC was more relevant for SJT performance in the low-incentive condition. While we screened out careless responders, it might be that differences in effort were slightly more influential in the low-incentive condition. In fact, a visual inspection of the scatterplots revealed a slight overrepresentation of individuals being low on the SJT and the ATIC in the low-incentive condition as compared to the high-incentive condition. However, excluding these participants did not substantially change our results. As another potential explanation, differences in the incentives given for SJT and ATIC performance may have led to a lower correlation in the high-incentive condition. In line with prior ATIC research (Ingold et al., 2015; Jansen et al., 2013; Melchers et al., 2009), we incentivized SJT but not ATIC performance in the high-stakes condition. While we also used an applicant framing in the low-incentive condition, no incentive was given for either SJT nor ATIC performance. This may have potentially created a higher symmetry in the motivational characteristics in the low- as compared to the high-incentive condition (i.e., low-SJT- and low-ATIC-incentive in the low-incentive condition vs. high-SJT and low-ATIC-incentive in the high-incentive condition). However, since we did not specifically formulate a hypothesis about the differences between high- and low-incentive conditions, we refrain from interpreting this finding in more detail and instead call for further research to understand the effects of incentives and simulated selection settings on correlates of SJT performance in general and, specifically, on the relationship between ATIC and SJT performance.

5 | GENERAL DISCUSSION

ATIC was found to be a relevant driver of performance in personnel selection simulations (Ingold et al., 2015; Jansen et al., 2013; König et al., 2007). However, research on the relevance of ATIC in SJTs was rare. The current research examined the relationship between ATIC and SJT performance across two studies. In Study 1, we administered items that tapped into 10 different constructs and came from five different SJTs with a behavioral-tendency instruction. In Study 2, we applied a full SJT on teamwork in a high- and low-incentive condition with a knowledge-instruction. We hypothesized that ATIC would predict SJT performance with a medium-sized positive effect. We

found that this hypothesis was partially supported; our data showed a significant but small effect of ATIC responses on SJT performance across items in Study 1 and in one of the two conditions in Study 2. Additional not-preregistered analyses revealed that the main driver behind differences in ATIC were due to the construct assessed in the SJT, not the individual.

5.1 | Theoretical implications

As a first finding, our results suggest that ATIC performance in SJTs depended substantially on the constructs that were addressed by the SJT items. This was further corroborated by our G-theory-based analysis (Woehr et al., 2012) showing that the largest proportion of variance was due to the constructs the items addressed. Particularly, the constructs of honesty/humility and employee integrity were much better detected by test-takers than other constructs such as teamwork, proactivity or specific team roles. This finding cannot be easily attributed to specific design choices made by test authors, since SJT tests accounted for much less variance in ATIC than did SJT constructs. This is further underlined by differences in ATIC scores *within* the HEXACO SJT (Oostrom et al., 2016), which were relatively high for honesty/humility but low for other HEXACO dimensions. We can only speculate that test-takers may be better able to pick up cues in SJTs for some constructs and less so for other constructs. Alternatively, some construct labels may be more familiar to test-takers than others and, therefore, easier to name correctly. Clearly, more research is needed to systematically identify relevant cues in SJTs, or more generally, to uncover the drivers behind the differences in ATIC performance across SJT items.

As a second finding and related to our hypothesis, we revealed that the included items in Study 1 as well as the entire Teamwork SJT in Study 2 yielded only a small relationship between ATIC and SJT performance. Mixed regression analysis attested that this was true on the individual level and on the item level. That is, individuals scoring higher on ATIC tended to show only slightly better SJT item responses (and vice versa). Moreover, items in which a better ATIC score was achieved did not yield better SJT item responses (and vice versa). The latter is particularly surprising considering that ATIC varied substantially across items—and much less across individuals—but still did not significantly explain SJT performance on the item level. So, one implication is that—contrary to findings in the realm of assessment centers and situational interviews (Ingold et al., 2015; Jansen et al., 2013; Kleinmann et al., 2011; König et al., 2007)—ATIC may be of little relevance in the herein applied set of SJT items (but also see Melchers & Hupp, 2017; Oostrom et al., 2016). In fact, the herein reported correlations were more similar to those found in studies in which ATIC was correlated with personality assessments (Barends & de Vries, 2023; Holtrop et al., 2021; König et al., 2006),⁹ thus potentially challenging the view of SJTs as simulations (e.g., Krumm et al., 2015).

Our results are in contrast to three previous studies, which revealed a stronger ATIC effect on SJT performance (Melchers &

Hupp, 2017; Oostrom et al., 2019; Wang et al., 2023). So, further and more detailed insights may be gained by comparing specifics of these studies. One difference between the herein reported Study 1 and the study by Melchers and Hupp (2017) was that the latter study used aggregated SJT responses and aggregated ATIC performances to obtain a correlation between SJT and ATIC performance. Aggregating scores across several constructs may minimize unique (but relevant) variance components of SJTs and instead increase shared ATIC variance. However, Study 2 used aggregated scores across items that addressed the same construct, but we still revealed lower correlations than the one reported by Melchers and Hupp. So, responding to items from the same SJT in direct sequence to each other may not help test-takers to gradually grasp an understanding of the measurement intentions, which they then potentially could have incorporated more and more into their response behavior. However, little is known about the processes that enable ATIC to unfold within simulations and further research is needed.

A notable difference between the current studies and the study by Oostrom et al. (2016) is that the latter used SJT items that came with a higher stimulus and response fidelity. That is, they presented video scenarios and had test-takers react directly to the actor in the video (i.e., they applied a behavioral response format). In line with previous findings from assessment center exercises or situational interviews (Ingold et al., 2015; Kleinmann et al., 2011), this open-ended SJT format showed a substantial relationship with ATIC. However, the vast majority of SJTs come in a closed response format, that is, they provide response options and are thus considered low-fidelity simulations. The herein reported results may, therefore, suggest that ATIC is more relevant in high(er)-fidelity simulations and less so in low(er)-fidelity simulations. A higher stimulus fidelity (e.g., video situations as compared to text situations) may provide test-takers with more cues on the situational demands (Naemi et al., 2016).

Furthermore, a higher response fidelity (e.g., constructed as compared to closed response formats) will enable test-takers to consider a lot of aspects when responding to assessments (including their thoughts on what the measurement intention of the assessment is). Closed response options in SJTs, on the other hand, have been found to be somewhat unrelated to the presented scenarios (Krumm et al., 2015; Schäpers et al., 2019), to elicit situation construal independently from the scenarios (Freudenstein et al., 2020), and to differ in their trait-relatedness (Schäpers et al., 2019). In sum, these aspects may make it hard for test-takers to apply ATIC. This is also in line with recent research on situational interviews, which revealed that including a dilemma resulted in an attenuated ATIC relevance on interview performance (Latham & Itzchakov, 2021). Potentially, response alternatives in SJTs function as a dilemma since they typically present contrary behavioral alternatives.

The closed response format may also explain differing results in the current studies and those reported by Wang et al. (2023). Notably, Wang et al. (2023) used a different way of assessing ATIC.

That is, they applied a closed-ended (multiple-choice) instrument to gauge ATIC scores. While scores obtained from closed-ended ATIC instruments were shown to be substantially related to assessment-center performance (Speer et al., 2014), their relation with SJT performance may be additionally driven by shared method variance due to the closed response formats. In the current studies, ATIC assessments and SJTs did not share the same response format and may therefore have resulted in lower correlations, as reported by Wang et al. (2023).

The current study followed prior ATIC research (e.g., Ingold et al., 2015; Kleinmann et al., 2011) and examined the relevance of ATIC in simulated personnel selection settings. That is, we provided a framing for the current studies by asking participants to imagine the questionnaire as being part of a selection process for their dream job. Contrary to prior studies, however, participants were acquired via a professional panel (prolific.co). Even though research suggests that (i) such panels generally provide good data quality (Peer et al., 2022), (ii) we checked for test motivation, and (iii) screened out careless responders, one might speculate that the personnel selection framing may have been less convincing (lower external validity in our study as compared to other (laboratory) studies), which might explain small relationships between ATIC and SJT responses. However, we also found that implementing a stronger performance incentive did not increase the relationship between ATIC and SJT performance. Ultimately, differences in study designs and in results between the currently available studies on ATIC in SJTs call for further research, which we suggest below.

5.2 | Practical implications

ATIC has been identified as an incremental predictor of job performance in prior research—above and beyond performance in a given selection method (Ingold et al., 2015; Jansen et al., 2013). Hence, SJT developers may be tempted to increase the predictive validity of SJTs by finding ways to design SJTs in a way that requires ATIC. However, the current results suggest that ATIC may not be a key driver of criterion-related validity in the herein applied SJTs. Moreover, our results suggest that particular choices made by test authors in developing low-fidelity SJTs will not necessarily translate into better or worse ATIC performances. That is, our results were relatively stable across different response instructions, traditional versus construct-oriented SJTs, and for SJT items versus entire SJTs. Comparing the current with a previous study by Oostrom et al. (2016) suggests that a higher ATIC saturation in SJTs might be achieved by increasing the fidelity of SJTs. In fact, a more realistic presentation of situations (e.g., in a video-based format) may provide test-takers with more cues about the measurement intentions, which can then be used to respond accordingly. However, we also like to emphasize that we employed several SJTs that yielded satisfactory construct validity and any attempt to increase the relevance of ATIC in such SJTs may be detrimental for their construct validity.

5.3 | Limitations and directions for future research

Several limitations must be addressed. First, although we selected typical representatives of the most common types of SJTs, our findings may not generalize to all SJTs, such as video SJTs or SJTs with a ranking format (Christian et al., 2010). Second, we relied on correlational designs. Thus, no causal inferences can be drawn from our results. Moreover, underlying third variables could have led to spurious correlations. For instance, one might speculate whether reading comprehension or social skills including the ability to grasp social situations might represent underlying drivers of the ATIC-SJT relationship. Given the generally low correlations in our studies, we are confident that our corresponding conclusions are not inflated due to spurious correlations.

We repeatedly pointed out that several notable differences exist between the currently available studies on the relevance of ATIC in SJTs (Melchers & Hupp, 2017; Oostrom et al., 2019; Wang et al., 2023). A final limitation of the current study thus is that we did not systematically incorporate these differences (e.g., in SJT item administration) and where thus not able to attest more firmly, how such differences affect ATIC responses and their relevance for SJT performance. Future studies should therefore (i) include SJT items that are presented in different fidelity (e.g., in a constructed as well as a closed response format) and (ii) present SJT items of the same construct consecutively as well as randomly mixed with other SJT items. Future research may also delve more deeply into the process of how ATIC unfolds over time within assessments. This might be done by examining the number of cues that need to be consistently available in an assessment until a hypothesis about the measurement intentions is formed, which is then applied to response behavior in the very same assessment.

6 | CONCLUSION

Across two consecutive studies, we sought to investigate the relevance of ATIC for SJT performance. On the one hand, we found a significant but small effect of ATIC responses on SJT performance across items in Study 1 and in one of the two conditions in Study 2. On the other hand, we also revealed that ATIC performance varied substantially across SJT items and constructs. We call for more research on the SJT design features that determine how well ATIC can be applied to SJT response behavior.

ACKNOWLEDGMENTS

This research was supported by a grant (KR 3457/2-2) from the German Research Foundation (DFG). Preliminary results of this research were presented at the 16th Conference of the Differential and Personality Psychology and Psychological Diagnostics (DPPD) section of the German Psychological Society and the 19th Conference of the Society of Industrial and Organizational Psychology (SIOP). The authors acknowledge the help of Johannes Horstmöller, Melanie Jacobsen, Gina Kriesmann, Nico Remmert, Julia Sauer, Chiara-Pauline Schreiber, Talea Stolte, and Tianqi Wang in coding

participant responses. Philipp Schäpers would like to thank the State of North Rhine-Westphalia's Ministry of Economic Affairs, Industry, Climate Action, and Energy as well as the Exzellenz Start-up Center. NRW program at the REACH-EUREGIO Start-Up Center for their kind support of our work. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework (OSF) at <https://osf.io/su9n4/>, reference number DOI 10.17605/OSF.IO/SU9N4.

ORCID

Nomi Reznik  <http://orcid.org/0000-0001-7294-5431>
 Stefan Krumm  <http://orcid.org/0000-0002-0840-0864> Jan-Philipp Freudenstein  <http://orcid.org/0000-0002-9029-5003> Pia Ingold  <http://orcid.org/0000-0002-6121-4227>
 Philipp Schäpers  <http://orcid.org/0000-0002-8270-5105>

ENDNOTES

¹ Our a-priori expectation was that situational judgment test (SJT) items could be seen by subject matter experts (SMEs) as possibly measuring constructs differing from the ones intended by the test authors. We thus preregistered to also use SMEs' alternative item construct ratings as another way of scoring ability to identify criteria (ATIC). However, SMEs identified the same constructs as intended by the test authors for all of the items.

² For two of the criteria (item clarity and content domain), initial interrater-reliabilities were insufficient (intraclass correlation coefficient ($ICC_{\text{clarity}} = .59$, $\kappa_{\text{domain}} = 0.48$). In response to this, raters were asked to discuss disagreements and reassess their ratings individually, which then resulted in sufficient to excellent interrater-reliabilities. $ICC > .90$, $\kappa > 0.70$).

³ We deviated slightly from this scoring logic for the HEXACO SJT, which consists of four response options that all present trait-related behavior but vary in the indicated standing on the trait. Also, no ineffective responses exist. We thus subtracted 7 points from the response reflective of the highest standing on the targeted trait. We subtracted 6, 5, or 4 points, respectively, for responses with gradually lower standing on the trait.

⁴ We preregistered two additional ways of scoring ATIC in SJTs. First, we were not sure in advance whether ATIC in SJTs could be differentially coded on a 4-point rating scale. Therefore, we also preregistered a binary rating scheme (0 = does not fit the construct, 1 = fits the construct). Second, we planned to apply an empirical scoring key, meaning that the most frequently given response is seen as the correct one (Bergmann et al., 2006). However, participant responses did not converge sufficiently to justify such a way of scoring; for most items, less than 20% of participant responses found the same (but different from the intended) construct.

⁵ Notably, this negative estimate is due to the distance scoring utilized in the SJT item scoring. Thus, this should be understood as a higher ATIC score entailing an SJT score that is less distant to the correct item solution.

⁶ We thank the editor and two anonymous reviewers for highlighting these issues.

⁷ Results remained stable when we imputed data points for these participants.

⁸ As a robustness check, we did the same analyses after eliminating the $N = 93$ participants with careless responses in the ATIC-part of the questionnaire. Results remained very similar, with test motivation still being higher in the high-incentive group ($M = 4.66$, $SD = 0.49$) than in the low-incentive group ($M = 4.29$, $SD = 0.69$, $t(350.55) = 6.2118$, $p < .01$), with a medium-sized effect ($d = 0.63$), and SJT performance also being significantly higher in the high-incentive group ($M = 10.31$, $SD = 3.95$) than in the low-incentive group ($M = 8.95$, $SD = 4.29$, $t(390.32) = 3.2909$, $p < .01$), with a small effect ($d = 0.33$).

⁹ We thank an anonymous reviewer for making us aware of this.

REFERENCES

- Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, 43(4), 695-716. <https://doi.org/10.1111/j.1744-6570.1990.tb00679.x>
- 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>
- Barends, A. J., & de Vries, R. E. (2023). Construct validity of a personality assessment game in a simulated selection situation and the moderating roles of the ability to identify criteria and dispositional insight. *International Journal of Selection and Assessment*, 31(1), 120-134. <https://doi.org/10.1111/ijsa.12404>
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H., Singmann, H., & Dai, B. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.
- Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment*, 13(3), 225-232. <https://doi.org/10.1111/j.1468-2389.2005.00319.x>
- 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). A 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>
- Cabrera, M. A. M., & 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>
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82(1), 143-159. <https://doi.org/10.1037/0021-9010.82.1.143>
- 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>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Freudenstein, J. P., R Emmert, N., Reznik, N., & Krumm, S. (2020). *English translation of the Teamwork Situational Judgment Test (SJT-TW)*. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). <https://doi.org/10.6102/zis285>
- Freudenstein, J. P., Schäpers, P., Reznik, N., Stolte, T., & Krumm, S. (2023). The influence of situational strength on the relation of personality and situational judgment test performance. *International Journal of Selection and Assessment*, 1-11. <https://doi.org/10.1111/ijsa.12444>
- 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*, 73(4), 669-700. <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, 203-208. <https://doi.org/10.1177/0963721416635552>
- Gatzka, T., & Volmer, J. (2017). *Situational judgement test for teamwork (SJT-TA)*. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). <https://doi.org/10.6102/zis249>
- Grand, J. A. (2020). A general response process theory for situational judgment tests. *Journal of Applied Psychology*, 105(8), 819-862. <https://doi.org/10.1037/apl0000468>
- Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, 26(3), 499-510. https://doi.org/10.1207/s15327906mbr2603_7
- Holtrop, D., Oostrom, J. K., Dunlop, P. D., & Runneboom, C. (2021). Predictors of faking behavior on personality inventories in selection: Do indicators of the ability and motivation to fake predict faking? *International Journal of Selection and Assessment*, 29(2), 185-202. <https://doi.org/10.1111/ijsa.12322>
- Hox, J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203852279>
- Ingold, P. V., Kleinmann, M., König, C. J., Melchers, K. G., & Van Iddekinge, C. H. (2015). Why do situational interviews predict job performance? The role of interviewees' ability to identify criteria. *Journal of Business and Psychology*, 30(2), 387-398. <https://doi.org/10.1007/s10869-014-9368-3>
- 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>
- 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>
- Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance*, 25(4), 273-302. <https://doi.org/10.1080/08959285.2012.703733>
- Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology*, 78(6), 988-993. <https://doi.org/10.1037/0021-9010.78.6.988>
- Kleinmann, M., & Ingold, P. V. (2019). Toward a better understanding of assessment centers: A conceptual review. *Annual Review of Organizational Psychology and Organizational Behavior*, 6, 349-372. <https://doi.org/10.1146/annurev-orgpsych-012218-014955>
- Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review*, 1(2), 128-146. <https://doi.org/10.1177/2041386610387000>
- König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. C. (2006). The relationship between the ability to identify evaluation criteria and integrity test scores. *Psychological Test and Assessment Modeling*, 48(3), 369-377.
- König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. C. (2007). Candidates' ability to identify criteria in non-transparent selection procedures: Evidence from an assessment center and a

- structured interview. *International Journal of Selection and Assessment*, 15(3), 283-292. <https://doi.org/10.1111/j.1468-2389.2007.00388.x>
- 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), 399-416. <https://doi.org/10.1037/a0037674>
- Latham, G. P., & Itzchakov, G. (2021). The effect of a dilemma on the relationship between ability to identify the criterion (ATIC) and scores on a validated situational interview. *Frontiers in Psychology*, 12, 674815. <https://doi.org/10.3389/fpsyg.2021.674815>
- 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>
- Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhard (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 279-300). Lawrence Erlbaum Associates Publishers.
- Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, 9(1), 3-22. <https://doi.org/10.1017/iop.2015.71>
- Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology*, 102(1), 43-66. <https://doi.org/10.1037/apl0000160>
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (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>
- 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., & Hupp, K. (2017, September 13-15). *Wie bedeutsam ist situative Urteilsfähigkeit für die Leistung in SJTs? [How relevant is situational judgment ability for performances in SJTs]* [Conference presentation]. Conference of the German Psychological Association, Section Work and Organisational Psychology, Dresden, Germany.
- Melchers, K. G., Klehe, U. C., Richter, G. M., Kleinmann, M., König, C. J., & Lievens, F. (2009). "I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance*, 22(4), 355-374. <https://doi.org/10.1080/08959280903120295>
- 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>
- 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>
- 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>
- Naemi, B., Martin-Raugh, M., & Kell, H. (2016). SJTs as measures of general domain knowledge for multimedia formats: Do actions speak louder than words? *Industrial and Organizational Psychology*, 9(1), 77-83. <https://doi.org/10.1017/iop.2015.121>
- 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>
- Oostrom, J. K., Melchers, K. G., Ingold, P. V., & Kleinmann, M. (2016). Why do situational interviews predict performance? Is it saying how you would behave or knowing how you should behave. *Journal of Business and Psychology*, 31, 279-291. <https://doi.org/10.1007/s10869-015-9410-0>
- Oostrom, J. K., de Vries, R. E., & de Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance*, 32(1), 1-29. <https://doi.org/10.1080/08959285.2018.1539856>
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), 1643-1662. <https://doi.org/10.3758/s13428-021-01694-3>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- 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>
- 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>
- 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., 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*, 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. (2020). The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology*, 105(8), 800-818. <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>
- Speer, A. B., Christiansen, N. D., Melchers, K. G., König, C. J., & Kleinmann, M. (2014). Establishing the cross-situational convergence of the ability to identify criteria: Consistency and prediction across similar and dissimilar assessment center exercises. *Human Performance*, 27(1), 44-60. <https://doi.org/10.1080/08959285.2013.854364>
- Thornton, G. C., & Cleveland, J. N. (1990). Developing managerial talent through simulation. *American Psychologist*, 45(2), 190-199. <https://doi.org/10.1037/0003-066X.45.2.190>
- Tiffin, P. A., Paton, L. W., O'Mara, D., MacCann, C., Lang, J. W. B., & Lievens, F. (2020). Situational judgement tests for selection: Traditional vs. construct-driven approaches. *Medical Education*, 54(2), 105-115. <https://doi.org/10.1111/medu.14011>

Wang, D., Oostrom, J. K., & Schollaert, E. (2023). The importance of situation evaluation and the ability to identify criteria in a construct-driven situational judgment test. *Personality and Individual Differences*, 208, 112182. <https://doi.org/10.1016/j.paid.2023.112182>

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>

Whetzel, D., Sullivan, T., & McCloy, R. (2020). Situational judgment tests: An overview of development practices and psychometric characteristics. *Personnel Assessment and Decisions*, 6(1), 1-16. <https://doi.org/10.25035/pad.2020.01.001>

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait-multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, 15(1), 134-161. <https://doi.org/10.1177/1094428111408616>

Wolcott, M. D., Lobczowski, N. G., Zeeman, J. M., & McLaughlin, J. E. (2021). Does the ability to identify the construct on an empathy situational judgment test relate to performance? Exploring a new concept in assessment. *Currents in Pharmacy Teaching and Learning*, 13(11), 1451-1456. <https://doi.org/10.1016/j.cptl.2021.09.003>

Wolcott, M. D., Lupton-Smith, C., Cox, W. C., & McLaughlin, J. E. (2019). A five-minute situational judgment test to assess empathy in first-year student pharmacists. *American Journal of Pharmaceutical Education*, 83(6), 6960. <https://doi.org/10.5688/ajpe6960>

SUPPORTING INFORMATION

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

How to cite this article: Reznik, N., Krumm, S., Freudenstein, J.-P., Heimann, A. L., Ingold, P., Schäpers, P., & Kleinmann, M. (2024). Does understanding what a test measures make a difference? On the relevance of the ability to identify criteria for situational judgment test performance. *International Journal of Selection and Assessment*, 32, 210-224. <https://doi.org/10.1111/ijsa.12458>

Appendix

Appendix A: Instructions for the Assessment of Ability to Identify Criteria (ATIC)

In answering each question, you may have had one or more assumptions about which **skill, trait, or competency (dimension(s))** your answer is most likely to assess.

For each question you answered, please write down in keyword form the **skill(s), trait(s), or competency(ies) (dimension(s))** that you assume you are looking for here. Also, please list behaviors that come to mind related to this skill, trait, or competency.

The laser printer of your department is located right next to the door of your office. The printer is used frequently and makes a lot of noise. You feel disturbed by the noise and find it hard to concentrate. Your colleagues appreciate the convenient position of the printer and are reluctant to have it moved. You could close the door to your office. However, you appreciate your department's practice of leaving doors open.

What would you do?

- a) I address the problem repeatedly and seek ways to have the printer moved.*
- a) I try to switch offices.*
- b) I try to get used to the noise of the printer.*
- c) I close the door of my office, when there is a lot of printing.*

<i>Dimension(s)</i>	<i>Behavior(s)</i>

Appendix B: Item Pool**Table B1***Items per Groups and Items Excluded from the Initial Item Pool*

Group	Test	Items
X-Set	Teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. 2020)	2, 5
	Personal Initiative SJT (Bledow & Frese, 2009)	6, 12
	Team Role Test (Mumford et al., 2008)	6, 8
	SJT for Employee Integrity (Becker, 2005)	2, 5
	HEXACO SJT (Oostrom, et al., 2019)	7 (H), 13 (H)
A-Set	Teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. 2020)	6, 7, 10
	Personal Initiative SJT (Bledow & Frese, 2009)	3, 10
	Team Role Test (Mumford et al., 2008)	1
	SJT for Employee Integrity (Becker, 2005)	8, 9, 10, 16, 19
	HEXACO SJT (Oostrom, et al., 2019)	16 (A), 18 (O), 19 (H), 21 (X)
B-Set	Teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. 2020)	11
	Personal Initiative SJT (Bledow & Frese, 2009)	1, 2, 5, 7, 9, 11
	Team Role Test (Mumford et al., 2008)	2, 3, 4, 5
	SJT for Employee Integrity (Becker, 2005)	1, 3, 4, 11
	HEXACO SJT (Oostrom, et al., 2019)	/
C-Set	Teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. 2020)	3, 4
	Personal Initiative SJT (Bledow & Frese, 2009)	4, 8
	Team Role Test (Mumford et al., 2008)	9, 7
	SJT for Employee Integrity (Becker, 2005)	17, 18
	HEXACO SJT (Oostrom, et al., 2019)	1 (H), 3 (X), 4 (A), 10 (A), 17 (C), 22 (A), 23 (C)
Excluded	Teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. 2020)	1, 8, 9, 12
	Personal Initiative SJT (Bledow & Frese, 2009)	/

Chapter 2: ATIC in SJTs

Team Role Test (Mumford et al., 2008)	10
SJT for Employee Integrity (Becker, 2005)	6, 7, 12, 13, 14, 15, 20
HEXACO SJT (Oostrom, et al., 2019)	2 (E), 5 (C), 6 (O), 8 (E), 9 (X), 11 (C), 12 (O), 14 (E), 15 (X), 20 (E), 24 (O)

Notes. All items were rated for containing a singular construct and providing clear, understandable information. 23 items were excluded for missing these criteria. Remaining items were randomly assigned to the three groups. Item numbers are corresponding to item numbers in the original SJTs. H = Honesty/Humility, E = Emotional Stability, X = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness for Experience.

Appendix C: Bivariate Correlations Among Ability to Identify Criteria (ATIC) Scores and Situational Judgment Test (SJT) Item Scores

Table C1

Means, Standard Deviations, and Bivariate Correlations Between Ability to Identify Criteria (ATIC) Scores and Situational Judgment Test (SJT) Item Scores in the A-Set

	ATIC Scores		SJT Item Scores														
	<i>M</i>	<i>S</i> <i>D</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Employee Integrity SJT (Item 10)	1.18	1.01	-	-	-.03	.10	.17	.10	.09	.15	.05	-.05	-.08	.13	.15	-.15	.05
2. Employee Integrity SJT (Item 16)	0.87	0.91	-	-	-.13	.07	.28*	.11	.13	.11	.07	-.01	.01	.09	.06	-.07	.12
3. Employee Integrity SJT (Item 19)	1.13	0.94	-	-	.20*	.18*	.08	.11	.11	.11	.04	-.04	.07	.06	.11	-.11	.06
4. Employee Integrity SJT (Item 8)	0.55	0.73	-	-	.06	.08	.01	.12	.05	.03	.22*	.18*	.06	.13	.07	-.12	.06
5. Employee Integrity SJT (Item 9)	0.87	0.78	-	-	.16	.07	.18*	.09	.01	.06	.00	.06	.00	.12	.04	.02	.05
6. HEXACO SJT (Item 16, A)	0.38	0.47	-	-	.02	.05	.15	.08	.22*	.08	.11	-.05	.11	.17	.08	.14	.01
7. HEXACO SJT (Item 18, O)	0.63	0.84	-	-	-.07	.01	.16	.03	.02	.01	.00	.16	.14	.44*	.02	-.10	.06
8. HEXACO SJT (Item 19, H)	1.53	1.29	-	-	.24**	.03	.08	.06	.03	.18*	.11	-.08	.06	.01	.12	-.06	.02
9. HEXACO SJT (Item 21, X)	0.40	0.80	-	-	-.08	.04	.16	.00	.14	.03	.20*	-.03	.13	.18*	.03	.16	.16
10. Personal Initiative SJT (Item 10)	0.36	0.75	-	-	-.04	.01	.21*	.09	.11	.11	.04	-.02	.15	.03	.11	-.11	.09
11. Personal Initiative SJT (Item 3)	0.32	0.73	-	-	-.02	.06	.21*	.06	.06	.09	.03	.01	.02	.14	.17	-.03	.08
12. Team Role SJT (Item 1)	0.53	0.66	-	-	.18*	.07	.05	.16	.19*	.11	.07	-.07	.10	.17	.02	-.04	.02
13. Teamwork SJT (Item 10)	0.67	0.96	-	-	-.16	.00	.13	.05	.02	.11	.05	.11	.03	.02	.17	.06	.02

Chapter 2: ATIC in SJTs

14. Teamwork SJT (Item 6)	0. 30	0.6 7	- .05	- .03	- .03	.0 4	.0 8	- .0	- .1	.0 2	- .1	- .0	- .0	- .0	- .0	.24 **	. 4
15. Teamwork SJT (Item 7)	0. 45	0.6 8	- .02	- .08	- .24**	.1 3	.1 6	- .0	.1 9*	.0 4	.0 3	.0 6	- .1	- .0	- .1	.02 0	. 5

Notes. Item numbers are presented to reflect the number assigned in the original SJTs: Teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. 2020), Personal Initiative SJT (Bledow & Frese, 2009), Team Role SJT (Mumford et al., 2008), Employee Integrity SJT (Becker, 2005), HEXACO SJT (Oostrom, et al., 2019). Inverted correlations are presented so that positive values reflect a positive relation between SJT item performance and ATIC responses. ATIC = Ability to identify criteria, SJT = Situational Judgment Test. H = Honesty/Humility, X = Extraversion, A = Agreeableness, O = Openness for Experience. * $p < .05$, ** $p < .01$.

Table C2

Means, Standard Deviations, and Bivariate Correlations Between Ability to Identify Criteria (ATIC) Scores and Situational Judgment Test (SJT) Item Scores in the B-Set

	ATIC Scores		SJT Item Scores														
	<i>M</i>	<i>S</i> <i>D</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Employee Integrity (Item 1)	0. 35	0.7 0	- .07	- .11	- .09	.16 .16	.0 .08	.0 .05	.0 .08	.0 .03	.0 .04	.0 .02	.1 .03	.0 .03	.1 .01	.0 .02	.0 .09
2. Employee Integrity (Item 11)	0. 65	0.8 6	- .04	.2 1*	.1 0	-.02 .00	.1 .07	.2 .2*	.1 .0	.0 .04	.1 .01	.0 .03	.0 .01	.1 .06	.1 .05	.0 .07	.1 .00
3. Employee Integrity (Item 3)	0. 53	0.7 6	- .03	.0 .07	.0 .01	.12 .12	.0 .09	.1 .08	.0 .09	.1 .05	.0 .01	.0 .03	.0 .07	.0 .01	.0 .01	.1 .01	.0 .07
4. Employee Integrity (Item 4)	0. 31	0.5 4	.17 .17	.0 .09	.0 .04	.00 .00	.0 .09	.0 .05	.0 .07	.0 .09	.0 .02	.0 .00	.1 .03	.1 .02	.0 .01	.1 .08	.0 .07
5. Personal Initiative SJT (Item 1)	0. 19	0.4 7	- .16	.0 .05	.0 .06	-.03 .00	.1 .01	.1 .08	.4* *	.1 .06	.0 .02	.0 .03	.0 .06	.0 .05	.2 0*	.0 .03	.1 .02
6. Personal Initiative SJT (Item 11)	0. 07	0.2 8	- .05	.0 .02	.0 .07	-.01 .00	.0 .00	.0 .05	.1 .05	.0 .05	.1 .04	.0 .07	.1 .00	.0 .06	.0 .06	.0 .04	.0 .00
7. Personal Initiative SJT (Item 2)	0. 30	0.7 4	-.1 .1	.1 .01	.1 .05	.16 .16	.1 .08*	.0 .04	.0 .01	.1 .02	.2 .0*	.1 .05	.0 .01	.0 .08	.1 .04	.0 .01	.0 .09
8. Personal Initiative SJT (Item 5)	0. 60	0.8 2	- .13	.1 .02	.0 .01	.02 .00	.0 .09	.1 .07	.0 .02	.0 .09	.0 .08	.0 .01	.0 .00	.1 .04	.0 .08	.1 .02	.1 .03
9. Personal Initiative SJT (Item 7)	0. 05	0.2 0	.08 .08	.1 .03	.0 .06	.07 .00	.0 .06	.0 .04	.1 .04	.0 .02	.0 .03	.1 .00	.1 .02	.0 .06	.0 .06	.0 .04	.0 .07
10. Personal Initiative SJT (Item 9)	0. 23	0.5 4	.19 *	.0 .03	.2 0*	-.06 .00	.1 .02	.1 .05	.0 .06	.0 .06	.0 .09	.0 .07	.0 .05	.1 .00	.1 .04	.0 .01	.2 .02*
11. Team Role SJT (Item 2)	0. 53	0.9 2	- .08	.1 .03	.0 .06	.26 **	.0 .06	.0 .07	.0 .06	.1 .00	.0 .08	.0 .06	.0 .00	.0 .04	.0 .00	.0 .04	.0 .05
12. Team Role SJT (Item 3)	0. 22	0.5 4	- .08	.0 .01	.1 .02	-.15 .00	.0 .01	.0 .04	.0 .01	.0 .06	.0 .02	.1 .04	.1 .02	.0 .09	.0 .03	.0 .04	.1 .08

Chapter 2: ATIC in SJTs

13. Team Role SJT (Item 4)	0.45	0.57	.10	-.05	-.04	.18	.16	.19*	.19*	.07	-.01	.12	.11	.02	-.01	.02	.07
14. Team Role SJT (Item 5)	0.54	0.70	-.10	-.17	-.07	.15	.17	-.03	.02	.00	.11	-.11	.12	-.02	.00	.07	-.04
15. Teamwork SJT (Item 11)	0.37	0.80	-.01	-.12	-.00	.14	.05	.11	.02	.11	.06	.05	-.00	.10	.05	.12	-.00

Notes. Item numbers are presented to reflect the number assigned in the original SJTs: Teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. 2020), Personal Initiative SJT (Bledow & Frese, 2009), Team Role SJT (Mumford et al., 2008), Employee Integrity SJT (Becker, 2005), HEXACO SJT (Oostrom, et al., 2019). Inverted correlations are presented so that positive values reflect a positive relation between SJT item performance and ATIC responses. ATIC = Ability to identify criteria, SJT = Situational Judgment Test. * $p < .05$, ** $p < .01$.

Table C3

Means, Standard Deviations, and Bivariate Correlations Between Ability to Identify Criteria

(ATIC) Scores and Situational Judgment Test (SJT) Item Scores in the C-Set

	ATIC Scores		SJT Item Scores														
	<i>M</i>	<i>S</i> <i>D</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Employee Integrity SJT (Item 17)	1.03	1.00	-	-.12	.12	.08	.02	.27*	.11	.11	.11	-.04	-.04	.01	-.06	-.06	.10
2. Employee Integrity SJT (Item 18)	0.12	0.41	.07	-.05	.03	-.03	.15	.08	.09	.09	.08	-.06	-.04	.05	.06	.11	.10
3. HEXACO SJT (Item 1, H)	2.48	0.86	-	-.08	.23*	.02	.06	.22*	.03	.03	.22*	-.14	-.03	.07	.09	-.06	.11
4. HEXACO SJT (Item 10, A)	0.19	0.39	-	-.04	.10	-.07	.10	.12	.11	.05	-.11	-.03	.00	.00	.04	.13	.11
5. HEXACO SJT (Item 17, C)	0.63	0.63	-	.23**	.08	.04	.03	.13	.00	.04	.04	-.08	.22*	.04	.04	.04	.04
6. HEXACO SJT (Item 22, A)	0.21	0.49	-	.05	.03	.09	.07	.06	.05	.05	.05	.05	.00	.19*	.05	.06	.11
7. HEXACO SJT (Item 23, C)	0.57	0.64	-	-.13	.29*	.13	.04	.30*	.07	.05	.21*	-.10	.07	.05	.02	.09	.03
8. HEXACO SJT (Item 3, X)	0.40	0.64	.08	.04	.06	-.04	.04	.03	.08	.01	.04	.07	.00	.01	.02	.07	.03
9. HEXACO SJT (Item 4, A)	0.48	0.58	-	-.02	.06	.02	.04	.06	.07	.06	.03	.06	.25*	.07	.01	.02	.07
10. Personal Initiative SJT (Item 4)	0.08	0.34	-	-.11	.06	.03	.15	.28*	.00	.05	.11	.03	.03	.05	.01	.01	.03
11. Personal Initiative SJT (Item 8)	0.37	0.80	-	-.03	.00	.03	.02	.00	.04	.03	.21*	.04	.00	.05	.09*	.11	.04
12. Team Role SJT (Item 7)	0.30	0.80	.11	-.11	.04	-.16	.03	.07	.03	.09	.08	.03	.08	.03	.03	.03	.07
13. Team Role SJT (Item 9)	0.10	0.37	.06	.05	.03	.03	.04	.11	.07	.03	.04	.05	.08	.08	.04	.04	.07
14. Teamwork SJT (Item 3)	0.19	0.63	-	.02	.22*	-.13	.12	.04	.01	.00	.11	.08*	.01	.07	.02	.04	.10
15. Teamwork SJT (Item 4)	0.49	0.78	-	-.09	.10	.25**	.02	.14	.01	.11	.02	.01	.12	.07	.08	.13	.12

Notes. Item numbers are presented to reflect the number assigned in the original SJTs: Teamwork SJT (Gatzka & Volmer, 2017; translated by Freudenstein et al. 2020), Personal Initiative SJT (Bledow & Frese, 2009), Team Role SJT (Mumford et al., 2008), Employee Integrity SJT (Becker, 2005), HEXACO SJT (Oostrom, et al., 2019). Inverted correlations are presented so that positive values reflect a positive relation between SJT item performance and ATIC

Chapter 2: ATIC in SJTs

responses. ATIC = Ability to identify criteria, SJT = Situational Judgment Test. H = Honesty/Humility, X = Extraversion, A = Agreeableness, C = Conscientiousness. * $p < .05$, ** $p < .01$.

Chapter 3

SJTs and Their Components

Greater Than the Sum of Its Parts? On the Relevance of Situational Judgement Tests'

Components for Construct and Criterion Validity

Nomi Reznik¹, Stefan Krumm¹, Jan-Philipp Freudenstein², Philipp Schäpers³

¹ Department of Education and Psychology, Freie Universität Berlin, Germany

² Hogrefe Verlagsgruppe GmbH, Group R & D, Germany

³ Department of Psychology and Sports, Universität Münster

This research was supported by a grant (KR 3457/2-2) from the German Research Foundation (DFG). Preliminary results of this research were presented at the 52. DGPs Congress in Hildesheim, Germany. We acknowledge the help of Anja Göritz for providing access to the recruiting infrastructure.

Abstract

This study investigates the unique and combined contributions of situational judgment test (SJT) components (situation descriptions, response options together and individually in randomized order, and open-format responses) to construct and criterion validity. Using a within-subjects design ($N = 121$), participants completed four item versions (response options only, randomized options with varied instructions, open-ended responses, and full items) across four waves. Results show that response options alone predicted full SJT items most consistently, aligning with recent critiques of the primacy of situational stems. While open-ended responses occasionally showed incremental validity, particularly for supervisor-rated job performance, effects were limited and inconsistent. Social desirability emerged as a robust predictor across conditions but there was little evidence that the relationship between SJT item versions and the respective construct's self-report measure was moderated by social desirability. The findings challenge traditional process models of SJT functioning and suggest that the response processes of goal formation and response evaluation may be more stable than context-sensitive reasoning models imply. Overall, the study highlights significant heterogeneity between SJT items in how item components relate to performance and calls for more differentiated theoretical and practical approaches to SJT development and validation.

Keywords: situational judgment test, construct validity, criterion validity, response format

Introduction

The emergence of Situational Judgment Tests (SJTs) in the 1990s marked a significant shift in the landscape of personnel selection. While their roots can be traced back to far earlier psychological assessments (Rosen, 1961), the modern form of SJTs is distinguished by its specific focus on workplace scenarios. Unlike traditional assessments that rely on abstract questions or purely cognitive evaluations, SJTs present applicants with realistic, job-related situations, offering a glimpse into how they might handle real-life challenges in the workplace. In doing so, they resemble a workplace simulation in text-based form (Motowidlo et al., 1990). In addition to their appeal as a work-related contextualized assessment, SJTs are more cost-effective than other contextualized assessments (e.g., assessment centers) and offer substantial predictive validity (Christian et al., 2010). As a result, SJTs have gained prominence in both practice and research.

The structure of SJT items typically involves a brief description of a situation, followed by a series of potential actions and/or judgments. Test-takers are asked to select the most appropriate response or rank the responses according to their effectiveness (McDaniel & Nguyen, 2001). Despite ample research, there remains a controversy around SJT functioning in general (Lievens & Motowidlo, 2016) and on the relevance of situation descriptions and response options in particular (Freudenstein et al., 2020; Schäpers et al., 2020; Schäpers et al., 2022; Schäpers, Lievens, et al., 2019; Schäpers, Mussel et al., 2019). Although several studies have examined the role of situation descriptions and response options for SJT functioning, these studies have examined varying combinations of SJT components; thus, naturally, arriving at different conclusions. While Rockstuhl et al. (2015) examined the relevance of situation versus response judgement in SJT items with an open response format, research by Krumm et al. (2015) addressed the relevance of situation descriptions in SJTs with a closed response format. A similar picture emerges for research on the underlying

processes in SJT response options. On the one hand, research by Leeds (2018) and Kaminski et al. (2019) emphasized characteristics of single response options, but Freudenstein et al. (2020), on the other hand, examined the relevance of the set of response options that is typically presented in SJTs. In light of these varying approaches to examining the underlying drivers of SJT functioning, research is needed that systematically varies all components of an SJT. In the current study, we will thus address this research gap by examining the different combinations of SJT components in a within-person design. In order to do this, we will examine the relevance of SJT item components for responding to the full SJT item as well as how SJT item components contribute to SJTs' construct- and criterion-related validity.

Theoretical Background

In 2016, Lievens and Motowidlo called for a reconceptualization of SJTs as measures of general domain knowledge. Their arguments were based on prior findings showing that a substantial proportion of SJT items can be solved without the situation descriptions being present (e.g., Krumm et al., 2015). However, several authors were opposed to this reconceptualization and presented arguments in favor of a more traditional view on SJTs (Fan et al., 2016; Melchers & Kleinmann, 2016). Notably, earlier studies arrived at different conclusions regarding SJT functioning. This is, among other underlying reasons, due to the fact that studies have taken different approaches in conceptualizing SJT functioning and focused on different SJT formats. Subsequently, we will disentangle these approaches and review research on the relevance of situation descriptions in SJTs with a closed response format, the relevance of situation descriptions in SJTs with an open response format, the relevance of response options as a set (i.e., all alternative responses to a situation description), and the relevance of single responses.

The Relevance of Situation Descriptions in SJTs With a Closed Response Format

For most of the time since their initial conceptualization, traditional SJTs (that is, SJTs with a closed response format, meaning a set of response options to choose from) have been considered to be situational measures, in that the situation description in the item stem has been considered to significantly drive the choice of response options in test-takers (Motowidlo et al., 1990). Several theoretical contributions that incorporate this view have been published since SJTs' revival in the 1990.

Ployhart (2006) put forward the predictor response process model, in which he proposed four steps in SJT answering: Comprehension, retrieval, judgement, and response selection. First, during the comprehension step, test-takers read, understand, and interpret the question or demand posed by the SJT item. Second, during the retrieval step, test-takers access information and knowledge regarding the given situation from their long-term memory, and third, during the judgement step, they form a judgement regarding the given situation on the basis of said stored knowledge. Finally, during the response selection step, test takers then check the response options offered by the SJT item and chose the one that best matches their own judgement.

Grand (2020) introduced the Situated Reasoning and Judgement Framework (SiRJ): according to the SiRJ, test-takers navigate through three distinct steps while answering SJT items. Within the first step, conditional reasoning, test-takers read and understand the item, thereby assessing the situation at hand and its specific demands. During the second step, similarity judgement, test-takers compare the identified situational demands with the set of response options. They thus juxtapose their own potential behavior that they assessed as fitting for the given situation with the behavioral options offered by the item. Following that, in the third step called preference accumulation, test-takers then chose the response option that fits best their self-generated behavior. Using computational modelling, Grand (2020) was

able to empirically demonstrate that the SiRJ might explain test takers' response behavior in SJT items.

In a similar vein, Martin-Raugh and Kell (2021) introduced the tripartite model of SJT responding, which also suggests three major cognitive processes: situation perception, in which test-takers perceive situational cues that are present in the SJT item and thus understand the presented situation; goal formulation, in which test-takers generate an intended goal or favorable outcome for the presented situation by evaluating the assessed situational cues and demands and aligning them with their individual values or expectations; and subsequently response evaluation, where test-takers then appraise the effectiveness of each response options for obtaining the goal formulated in the previous step.

In all three of these process models, the authors conclude that test-takers base their understanding of the item and its demands on the information given in the situation description, before examining the response options and choosing a response. This conceptualization thus places particular importance on the role of situation descriptions as an important driver of SJT response processes, and in the case of the SiRJ, the computational modelling even provides empirical evidence on this.

The relevance of situation descriptions in SJTs, however, was placed under heavy scrutiny following the findings of Krumm et al. (2015): Their research suggested that many SJT items could be effectively solved even without the inclusion of detailed situational context (that is, without the situation description). This was later backed and furthered by Schäpers, Mussel et al. (2019) who found that omitting situation descriptions did not significantly impact either test scores or participant perceptions of SJTs, and that construct saturation of both personality as well as ability remained similar regardless of whether situation descriptions were present. Additionally, Jackson et al. (2017) found that most variance in SJT responses was due to individual effects and not situation effects. These

findings challenged the traditional view on SJTs, which holds that the context provided by these descriptions is the crucial driver for test-taker's judgment (Campion & Ployhart, 2013). In summarizing prior research on situation descriptions, we conclude that ample theoretical accounts of SJT functioning emphasize their relevance. Empirical evidence, however, has so far provided support for both views, that is, for and against the relevance of situation descriptions.

The Relevance of Response Options as a Set

As a reaction to the findings of Krumm et al. (2015), other researchers (Fan et al., 2016; Melchers & Kleinmann, 2016) proposed that SJTs retain their situational essence even without explicit situation descriptions. According to their view, even when the situational descriptions are absent, test-takers can infer the relevant context and make informed decisions based on the cues provided by the response options (Melchers & Kleinmann, 2016). Indeed, the reviewed research in the previous section did not employ research designs which ruled out that situation descriptions may be construed from the set of response options.

Only few studies are so far available that provide specific insights on how response sets in SJTs enable the construal of situations. As one of these studies, Freudenstein et al. (2020) demonstrated that the situation construal of SJT items, based solely on response options, significantly influenced participant responses. The authors applied several sets of SJT items with and without situation descriptions, and asked test-takers to additionally assess their situation construal, measured by the situational-8 DIAMONDS (Rauthmann & Sherman, 2015). They found that situation construal was significantly related to SJT performance, which was the case even when situation descriptions were omitted. Further, indirect evidence for the relevance of response options can be derived from a study by Schäpers, Lievens et al. (2019). These authors found that SJT scores were higher when the formats of situation descriptions and response options matched. That is, when video

descriptions were followed by video response options or text-based descriptions were followed by written response options, average SJT performance was higher than in a mixed design with incongruent formats. It may be argued that response options that were presented in a different format than the situation descriptions were less conclusive for the situation at hand, thus pointing to the relevance of response options as a source of (context) information in SJTs.

We conclude that while not many studies have so far examined the situation construal made on the basis of SJT response options, the few studies who did so demonstrated that test-takers can infer situational context from response options alone. This suggests that response options play a crucial role in providing contextual information and shaping situation construal in SJTs.

The Relevance of Individual Response Options

When assuming that, generally, SJT response options are in fact offering rich amounts of contextual information, the question remains what processes are then at work when test-takers understand these response options, and what processes then lead to them choosing a specific one. The relevance of response options might be investigated by looking into the aforementioned process models (Grand, 2020; Martin-Raugh & Kell, 2021; Ployhart, 2006). All of these models specifically address the process of how test-takers choose response options in a similar way: test-takers understand the situational demands, generate a behavioral response to these demands, and then carefully understand and interpret the information in each individual response options before vetting all response options against their own, self-generated response; finally they chose the one response options whose characteristics best match those of the self-generated behavior.

Fittingly, empirical research has been conducted to examine characteristics of SJT response options. The work of Leeds (2018) and Kaminski et al. (2019) has highlighted the

importance of the individual response options in SJTs, particularly focusing on the concept of response options' social desirability. Leeds (2018) introduced the Theory of Cognitive Acuity (TCA) for SJT development, treating response options as signals indicating the correctness or effectiveness and emphasizing the importance of test-takers' sensitivity to these signals. This approach suggests that individuals' ability to discern the most appropriate responses is linked to their sensitivity to the subtleties of correctness embedded within the response options. Adding to this, Kaminski et al. (2019) explored the influence of social desirability in SJT response options. Social desirability refers to the tendency of individuals to respond in a manner that they think is socially acceptable or portrays themselves in a positive light, and has been found to play an important role in personnel selection, as well as in most psychological assessment measures (McFarland & Ryan, 2006). This means that for SJTs and other measurement methods, social desirability could impact the reliability and validity of the test results, as individuals may not provide honest and genuine responses, but rather ones that they believe the test administrators would prefer. Indeed, Kaminski et al. (2019) found that both the social desirability as well as the general plausibility of SJT response options (as rated by subject matter experts) significantly affects test-taker responses across different SJT formats (that is, with a standard instruction as intended by the test authors, with a standard instruction but without the situation description, and with a faking-good instruction). This indicates that the attractiveness of response options, in terms of social desirability, plays a critical role in how individuals respond to SJTs. Moreover, we argue that research on the relevance of SJT components should also be conducted on the level of single response options (and not just sets of response options) as single response options may give away information about the situational demands.

The Relevance of Situation Descriptions in SJTs With an Open Response Format

With all these insights on the functioning of the different SJT parts in closed-response format SJTs, questions remain what processes are at work when working on SJTs in which only the situation and no response options are presented at all, thus reflecting an open response format. To date, we are aware of only one study that examined the relevance of situations in SJTs with an open response format (Rockstuhl et al., 2015). Specifically, these authors made the distinction between situational judgement (measured via asking test-takers what they believed the individuals in the presented situations were thinking, feeling, and intending with their actions) and response judgement (measure by asking test-takers what they would do in the presented situation). Their results showed that open-format judgments of situations were not only moderately correlated with judgement of responses, but also offered incremental validity in predicting task and contextual performance beyond judgment of responses, thereby underscoring the value of situational context in SJTs. Thus, contrasting the view of reduced importance of situation descriptions (e.g., Krumm et al., 2015; Schäpers. Mussel et al., 2019), Rockstuhl et al. (2015) provided evidence of their relevance and concluded that situational judgment is indeed an important process in responding to SJT items. In other words, results from open response SJTs support the traditional view of SJTs, which emphasizes the importance of situation descriptions in SJTs. Their research suggests that these descriptions do more than just set the scene; at least in open response format SJTs.

Conclusion

The current debate in the field of SJTs as well as the mixed findings highlight the multifaceted nature of these assessments. Thus, a comprehensive understanding of SJTs requires considering all their components (the situation descriptions, the set of response options, and the characteristics of each response option) and combinations thereof. Currently, there is a need for more empirical studies to understand how these different elements interact and contribute to the overall effectiveness of SJTs

To address this, we administered four distinct versions of the same SJT item in a within-subjects design, exploring the individual and combined effects of situation descriptions, response options, and their permutations. In doing so, we bring together the previously separated design choices by Krumm et al. (2015), Leeds (2018), Kaminski et al. (2019), and Rockstuhl et al. (2015), thereby aiming to provide a comprehensive understanding of the elements that contribute to the effectiveness of SJTs. The four SJT-variations we applied were (i) SJT items without situation descriptions (following the works of Krumm et al., 2015; Schäpers, Mussel et al., 2019), (ii) only the response options individually and their respective properties (following the works of Kaminski et al., 2019; Leeds, 2018), (iii) the situation descriptions alone, that is, in an open-response format (following the works of Rockstuhl et al., 2015), and (iv) the full item with situation description and response options. As such combinations of SJT elements have not yet been examined before in a within-subjects design, we follow a mostly exploratory approach and propose the following research questions and hypotheses (as preregistered; see https://osf.io/nt9w4/?view_only=b5e85bd8237a479ca2c5ae5ca752b267):

RQ1: To what extent do item versions i, ii and iii predict the full item version (iv)? That is, what is the relative contribution of versions i to iii in predicting the full item?

RQ2a: To what extent do item versions i, ii, iii, and iv predict self-reports of the targeted construct?

RQ2b: Will there be a significant incremental contribution of versions i, ii, and iii over iv and vice versa in predicting self-reports of the targeted construct?

As there is great similarity between SJT version ii and traditional self-report questionnaires, we follow the logic of Campbell and Fiske (1959) and hypothesize:

H1: There will be an incremental contribution of SJT version ii above and beyond version iv in predicting self-reports of the targeted construct.

As delineated in the preceding discussion, Kaminski et al. (2019) demonstrated that the social desirability ratings of response options within Situational Judgment Tests (SJTs) account for variations in item performance. Similarly, self-reported personality assessments can be subject to bias stemming from socially desirable responding (Bradley & Hauenstein, 2006). Consequently, it is assumed that an individual's propensity to engage in socially desirable behavior, hereafter termed 'trait social desirability,' could serve as a previously unexplored factor influencing the correlation between SJT responses and the corresponding measures of self-reported constructs. Accordingly, we propose the following research question:

RQ3: Does trait social desirability moderate the relationship between SJT item versions and the respective construct's self-report measure?

Rockstuhl et al. (2015) established that situation judgement in Situational Judgment Test (SJT) items significantly enhances the predictive validity for job-related criteria beyond the predictive capability of SJT response judgement. In further examining the construct- and criterion-related validity, we also phrased the following research questions.

RQ4a: To what extent do item versions i, ii, iii, and iv predict self- and supervisor-rated job performance?

RQ4b: Will there be a significant incremental contribution of versions i, ii, and iii over iv and vice versa in predicting self- and supervisor-rated job performance?

Finally, we phrased the following hypothesis based on prior findings that open response SJT items incrementally predicted job-related criteria above and beyond SJT items presented in a closed format:

H2: There will be an incremental contribution of version iii above and beyond iv in predicting self- and supervisor-rated job performance.

Method

Participants

Participants were recruited via a German online panel (<https://www.wisopanel.net/>) and compensated with 22€ in total for completing all sessions and providing supervisor feedback. Initially, we contacted 400 participants for session A and then recontacted these participants for each additional session. After excluding 26 careless responders, that is, participants who failed more than one out of two attention checks per session, a final $N = 121$ (43.8% female, $M_{Age} = 46.98$, $SD_{Age} = 11.85$) completed all four waves. Most of the participants were employed full-time (73%), with a mean work duration of 11.5 years. Most of them were employed in either service industry jobs (35%) or production (14%). The total sample size falls short of our a priori power analysis that pointed to a sample of 144 participants for our moderation analyses (https://osf.io/nt9w4/?view_only=b5e85bd8237a479ca2c5ae5ca752b267). Even though providing supervisor feedback was heavily incentivized, supervisor response rate was low, which was exacerbated by technical problems during data collection and resulted in a final $N_{supervisor} = 61$. An additional 50 new participants were recruited for sessions C and D each, to compare performance across all session types and to identify any sequence or memory effects comprehensively. These participants were only recruited to control for sequence effects and were incentivized with 3€ but were not included in our main analyses.

Design, Materials, and Procedure

A within-subjects design was employed in this study. Participants completed four test sessions containing eight SJT items each in different versions i to iv as delineated above (for descriptions of SJT items in each of these four versions, see Appendix A). The original eight items that served as a basis our item manipulations were sourced from two published SJTs, the Personal Initiative SJT (Bledow & Frese, 2009) and an SJT on facets of the Five Factor

personality model (Mussel et al., 2018), thus facilitating that our findings generalize across SJTs, from different domains. For the personality SJT, the focus was on the conscientiousness domain due to its known predictive validity for job performance (Salgado & Tauritz., 2014). We administered items from a previously validated shortened version of the conscientiousness SJT, which demonstrated the highest item-total correlations (Schäpers et al., 2020). For the Personal Initiative SJT, prior research revealed that some items can be solved regardless of the situation description being available, while this was not possible for other items (Freudenstein et al., 2020). To account for this, we included items from both of these categories.

All close-ended SJT items were scored using distance scoring: The distance from the correct response was measured on a 7-point Likert scale, with points deducted based on the correctness of the response: -7 for correct answers, -4 for distractors, and -1 for incorrect responses (for in-depth explanation of distance scoring, see Reznik et al., 2023). Open responses were coded with respect to their level of trait manifestation by two independent coders. Specifically, coders used the original closed-responses to rate open responses from participants. ICCs were all good to excellent, ranging from .78 to .94, (see LeBreton & Senter, 2008), thus allowing for averaging of codings. Test sessions were designated as sessions A, B, C, and D, each spaced 2-3 weeks apart. To control for sequence effects in response, participants were randomly assigned to one of two sequences (A-B-C-D or B-A-C-D). Sessions C and D always preceded A and B to mitigate memory effects.

To minimize participant burden, each session was designed to be as brief as possible while still presenting a variety of SJT items. During session A, participants were first queried about their ability to provide supervisor performance ratings. Only those affirming proceeded with the survey. Session A then proceeded with SJT items administered in an open-response format (version iii). In session B, participants were presented with a pseudo-randomized list

of all response options (version ii) that were taken from all the original SJT items. We administered pseudo-randomized response options in three different instructions to each participant (see Ployhart & Ehrhart, 2003): a behavioral tendency instruction (i.e., “I would behave like this”), an effectiveness instruction (i.e., “I find this behavior to be effective”), and a social desirability instruction (i.e., “I find this behavior to be socially desirable”). Each response option was rated on a 7-point scale. In session C, SJT items without item stems (version i) were administered. Finally, in session D, the complete, unmanipulated SJT items (version iv) were presented. Participants also completed self-ratings of the same scales as their supervisors (IRB, OCB, conscientiousness, personal initiative, and trait social desirability) in session D. Supervisors were contacted by the participants via email and were asked to rate job performance, operationalized by in-role behavior (IRB; Williams & Anderson, 1991), and organizational citizenship behavior (OCB; Williams & Anderson, 1991), of their employee.

Analytic Strategies

The data were analyzed using hierarchical regression models, focusing on item-level analyses to determine the impact of each SJT version in predicting full item versions and self- and supervisor reports. Relative weights analyses are included in the appendix, but did not provide information above and beyond regression analyses, as we initially assumed more predictors would become significant. Moderated regression was used to examine interaction effects between item versions and trait social desirability. All analyses were conducted in RStudio using packages *psych* (Revelle, 2025), *stats* (Base R) and *pwr* (Champely et al., 2020).

Results

Preliminary Analyses

Sequence Effects

Welch's *t*-tests (Appendix B) were conducted to examine potential differences between the two groups with different wave orders, as well as between regular and additional participants in sessions C and D. No significant differences were observed, even after applying Bonferroni corrections for multiple comparisons. These findings suggest that the time between sessions was sufficient to prevent memory, fatigue, or learning effects due to the repetitive nature of the study design.

Descriptive Analyses

Reliability estimates (see Tables 1 and 2) were satisfactory for all scales but SJT item version iii (open-ended), which showed low internal consistency. Correlation analyses (see Table 3) did reveal some unexpected negative relationships. For example, version iii (open-ended) showed a significant negative correlation with supervisor-rated OCBI for personal initiative ($r = -.47, p < .001$), and version i (without stems) showed a significant negative correlation with supervisor-rated OCBI for conscientiousness ($r = -.25, p < .05$). Furthermore, trait social desirability correlated only weakly with version ii under the social desirability instruction ($r = .10, ns$).

Table 1

Descriptive Item Statistics and Scale Reliability for SJT Item Versions i, iii, and iv

SJT Item	Version i (without stems)				Version iii (open ended items)				Version iv (full item)			
	<i>M</i>	<i>SD</i>	<i>r</i> _{it}	ω	<i>M</i>	<i>SD</i>	<i>r</i> _{it}	ω	<i>M</i>	<i>SD</i>	<i>r</i> _{it}	ω
Personal Initiative	2.48	.54		.66	.46	.6		.38	2.39	.57		.67
PI7	2.26	.7	.64		.35	.3	.82		2.42	.76	.73	
PI10	2.78	.84	.61		.37	.61	.54		2.7	.98	.76	
PI1	2.89	.74	.80		.52	.63	.25		2.66	.88	.69	
PI9	1.99	.85	.67		.6	.92	.71		1.78	.78	.48	
Conscientiousness	2.11	.54		.69	.44	.89		.1	2.32	.56		.72
C15	1.92	.74	.72		.43	.89	.28		2.36	.92	.67	

Chapter 3: SJTs and Their Components

C20	2.07	.79	.72	.28	.98	.58	2.2	.73	.72
C18	1.98	.64	.60	.45	.86	.38	2.08	.63	.6
C12	2.46	.92	.72	.59	.8	.89	2.64	.98	.71

Note. $N = 121$.

Table 2

Descriptive Item Statistics and Scale Reliability for SJT Item Version ii

SJT Item	Version ii (behavioral tendency)				Version ii (knowledge)				Version iv (social desirability)			
	<i>M</i>	<i>SD</i>	r_{it}	ω	<i>M</i>	<i>SD</i>	r_{it}	ω	<i>M</i>	<i>SD</i>	r_{it}	ω
Personal Initiative	2.54	.48		.57	2.38	.49		.57	2.5	.42		.48
PI7	2.24	.71	.63		2.4	.74	.68		2.38	.76	.66	
PI10	2.81	.78	.66		2.35	.89	.44		2.6	.71	.59	
PI1	2.99	.65	.64		3.08	.63	.69		3.04	.66	.46	
PI9	2.13	.89	.61		1.86	.87	.66		1.97	.75	.63	
Conscientiousness	2.21	.52		.67	2.08	.61		.78	1.95	.5		.63
C15	1.99	.69	.64		1.78	.69	.55		1.72	.74	.52	
C20	2.1	.77	.75		2.05	.83	.75		1.98	.7	.62	
C18	2.23	.7	.61		1.96	.68	.78		1.96	.69	.77	
C12	2.53	.87	.73		2.52	1.05	.85		2.14	.86	.75	

Note. $N = 121$.

Table 3*Descriptive Statistics and Bivariate Correlations for Self- and Supervisor Reports with SJT Item Versions*

Variable	<i>M</i>	<i>SD</i>	ω	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Self-reports																		
1. Personal Initiative	5.08	1.07	.91	—														
2. Conscientiousness	5.42	1.06	.8	.63***	—													
3. Job Performance (IRB)	6.14	1.03	.95	.38***	.57***	—												
4. Job Performance (OCBI)	5.37	1.17	.95	.43***	.3***	.46***	—											
5. Trait Social Desirability (TSD)	4.37	.68	.79	.11	-.06	.11	.26**	—										
Supervisor reports																		
6. Personal Initiative	5.62	1.17	.95	.64***	.43**	.24	.56***	-.02	—									
7. Conscientiousness	6.12	.98	.87	.46***	.53***	.49***	.39**	-.08	.71***	—								
8. Job Performance (IRB)	6.36	.86	.94	.31*	.38**	.47***	.38**	-.07	.5***	.8***	—							
9. Job Performance (OCBI)	5.87	1.13	.95	.51***	.41**	.37**	.68***	.12	.81***	.64**	.54***	—						
SJT versions																		
Personal Initiative																		
10. Version i (without stems)	2.48	.54	.66	-.13	-.02	.21*	-.08	.13	-.23	.09	.14	-.23	—					
11. Version ii (beh. tendency)	2.54	.48	.57	-.08	.01	.23**	.02	.08	-.06	.26*	.29*	-.09	.58**	—				
12. Version ii (knowledge)	2.38	.49	.58	-.08	.15	-.03	-.2*	-.004	-.24	-.01	-.03	-.36*	.47**	.5*	—			
13. Version ii (desirability)	2.5	.42	.48	.1	.11	.16	.10	.10	-.06	-.02	.14	-.11	.23**	.42**	.5**	—		
14. Version iii (open ended)	.46	.6	.38	-.26	.02	.13	-.3	.17	-.47	-.39	.06	-.47	.25	.01	.07	.05	—	
15. Version iv (full item)	2.39	.57	.67	-.21*	.2*	.1	-.12	-.02	-.14	.12	-.03	-.28*	.58**	.46**	.28**	.14	-.05	—
SJT versions																		
Conscientiousness																		

Chapter 3: SJTs and Their Components

Variable	<i>M</i>	<i>SD</i>	ω	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
10. Version i (without stems)	2.11	.54	.69	-.17	-.02	.05	-.13	-.06	-.35*	-.22	-.04	-.25	—					
11. Version ii (beh. tendency)	2.21	.52	.67	-.16	.01	.11	-.07	-.18*	-.26	-.13	-.06	-.17	.48**	—				
12. Version ii (knowledge)	2.08	.61	.78	.07	.15	.09	-.06	-.13	.09	.26	.25	.13	.37**	.63**	—			
13. Version ii (desirability)	1.95	.5	.63	-.08	.11	.03	-.14	-.16	-.15	-.20	-.26	-.27*	.33**	.53**	.5**	—		
14. Version iii (open ended)	.44	.89	.1	-.02	.02	.04	.05	-.03	-.04	-.17	-.06	-.03	.08	-.04	-.08	-.1	—	
15. Version iv (full item)	2.32	.56	.72	-.05	.2*	.35**	.04	-.1	-.19	.14	.23	-.03	.52**	.55**	.51**	.3**	-.09	—

Note. $N_{\text{self}} = 121$. $N_{\text{supervisor}} = 61$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Main Analyses

RQ1: Predicting the Full Item (Version iv)

To investigate the relative contribution of SJT item components, hierarchical regression analyses were conducted for each item. In Step 1, version i (response options without stems) was entered. In Step 2, version ii (ratings of randomized response options under behavioral tendency, knowledge, and social desirability instructions) was added. In Step 3, version iii (open-ended responses to the situation description) was included. Adjusted R^2 values, changes in explained variance (ΔR^2), and significance tests (ΔF) for each step are presented in Table 4.

Across most items, version i emerged as the strongest individual predictor of full item responses (version iv). The addition of version ii variables resulted in small to moderate increases in explained variance, with significant ΔR^2 observed in several items (e.g., PI7: $\Delta R^2 = .05, p = .059$; C15: $\Delta R^2 = .13, p < .01$; C12: $\Delta R^2 = .19, p < .01$). The open-ended version iii did not provide significant incremental validity beyond the other components for any item. In all models, adjusted R^2 increased only marginally or not at all after version iii was entered, and ΔF values were consistently non-significant (all $p > .05$).

Relative weights analyses (see Appendix C) supported these findings, showing that version i accounted for the largest proportion of explained variance in nearly all items. The patterns of predictor importance were consistent across both methods, with no additional insights gained from the relative weights approach beyond those obtained from the hierarchical regression.

Table 4

Hierarchical Regression Analyses with SJT Item Versions i to iii Predicting Item Version iv

Regression	Personal Initiative SJT (version iv; full item)				Personality SJT (version iv; full item)			
	PI7	PI10	PI1	PI9	C15	C20	C18	C12
Step 1: Version i (without stems)								
β	.61***	.35	.44**	.51***	.28*	.36***	.37***	.36**
Adj. R^2	.3	.07	.15	.29	.04	.14	.12	.09
Step 2: Version ii (random responses)								
β (behavioral tendency)	.21*	.07	.26	-.01	.35*	.16	.03	.06
β (knowledge)	.05	-.14	-.23	-.04	.26	-.028	.11	.43***
β (desirability)	.09	.008	.01	.16	.007	.029	-.02	-.22
Adj. R^2	.33	.006	.15	.27	.15	.14	.11	.26
ΔR^2	.05	.014	.05	.02	.13**	.03	.02	.19**
ΔF	2.57	.17	1.06	.73	5.37	1.16	.69	7.34
Step 3: Version iii (open ended)								
β	-.002	-.09	.01	.01	-.02	.003	-.01	.001
Adj. R^2	.33	.02	.14	.27	.15	.138	.11	.25
ΔR^2	.00	.00	.00	.01	.01	.00	.00	.00
ΔF	.03	.14	.5	1.54	1.39	.04	.79	3.58

Note. $N_{\text{self}} = 121$, $N_{\text{supervisor}} = 61$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

RQ2: Prediction of Self-Reported Constructs

RQ2a. Hierarchical regressions assessed how well item versions predicted self-reported personal initiative and conscientiousness (see Appendix C for detailed information of all coefficients). Overall, predictive patterns were heterogeneous and varied between items. For personal initiative, full model fit varied from negligible to moderate across items (e.g., for item PI9: adj. $R^2 = .19$, for item C15: adj. $R^2 = .05$). Several version ii indicators were significant, such as for item PI1 behavioral ($\beta = .44$, $p < .05$), suggesting item-specific

relationships with self-reported constructs. For conscientiousness, explained variance remained generally low ($\text{adj. } R^2 \leq .09$). Still, version ii showed some significance (e.g., a negative relationship for item C15 behavioral: $\beta = -.34, p < .05$), while version iii did not significantly predict self-reports in any item.

RQ2b. Incremental validity was evaluated in both directions. When versions i–iii were added to version iv, explained variance did not improve significantly (ΔR^2 s between .00 and .15, *n.s.*). Significant contributions stemmed mostly from version ii predictors (for example for item C12 knowledge: $\beta = .31, p < .05$). When version iv was added to models already including i–iii, explained variance improved in some items (items PI10, PI9, C20 with ΔR^2 ranging from .06 ($p < .05$) to .14 ($p < .01$), indicating that the full item added some additional, incremental value beyond its components.

Taken together, these results indicate that while some item versions (mostly version ii) were predictive of for the corresponding constructs, the full item contributed incremental variance in some cases, supporting its potential added value beyond its individual components.

H1: Incremental Contribution of Item Version ii Above and Beyond Version iv in Predicting Self-Reports of Corresponding Constructs

To test H1, we conducted hierarchical regressions to examine whether version ii explained incremental variance in self-reported personal initiative and conscientiousness beyond the full item (see Table 5). Results showed significant ΔR^2 values for only two items (e.g., PI1: $\Delta R^2 = .07, p < .05$; C15: $\Delta R^2 = .09, p < .05$), with significant predictors stemming from version ii, particularly social desirability (e.g., PI1: $\beta = .36, p < .05$, C15 behavioral: $\beta = -.53, p < .01$). However, for the other six items, no incremental variance was explained by version ii beyond the full item, which is why we consider our H1 not supported.

Table 5

Incremental Contribution of Item Version ii Beyond the Full Item in Predicting Corresponding Self-Reported Constructs

Regression	Personal Initiative				Conscientiousness			
	PI7	PI10	PI1	PI9	C15	C20	C18	C12
Step 1: Version iv (full item)								
β	-.13	-.14	-.02	-.48**	.18	.33*	.22	.13
Adj. R^2	.00	.01	.00	.12	.02	.04	.01	.01
Step 2: Version iv + version ii (random responses)								
β (full item)	-.12	-.16	-.07	-.41**	.26*	.36**	.18	.02
β (behavioral tendency)	-.14	.11	.19	-.07	-.53**	-.23	.12	-.07
β (knowledge)	-.03	-.06	-.20	-.06	.19	.07	.07	.23
β (desirability)	.35*	.08	.36*	-.11	.07	.21	-.01	-.02
Adj. R^2	.04	-.01	.03	.12	.08	.05	.00	.01
ΔR^2	.06	.00	.07*	.02	.09*	.03	.01	.03
ΔF	2.46	.33	2.77	.84	3.80	1.39	.42	1.13

Note. $N = 121$.

* $p < 0.05$. ** $p < 0.01$. *** $p < .001$.

RQ3: Moderation Effects of Trait Social Desirability

Moderated regression analyses were conducted to examine whether trait social desirability (TSD) moderated the relationship between item versions and the self-reported corresponding traits personal initiative and conscientiousness. Notably, post-hoc power analysis revealed that with a sample size of $N = 121$, seven predictors, and a significance level of $\alpha = .05$, the study had a power of 0.45 to detect a small effect ($f^2 = 0.08$). For transparency's sake we include all results but advise to interpret with caution. Across all tested interactions, few reached significance (see Table 6). For personal initiative, one significant interaction emerged between TSD and the full item in PI7 ($\beta = -.58, p < .05$),

indicating that higher TSD was associated with a reduced relationship between the item and self-reported personal initiative. For conscientiousness, significant interactions were found for item C15: between TSD and version ii ($\beta_{iidesirability} = .65, p < .01$; $\beta_{iiknowledge} = .63, p < .05$), suggesting that higher TSD strengthened the association between these item versions and self-reported conscientiousness. No other interaction terms were significant.

Table 6

Moderated Regression Analyses with SJT Item Versions i to vi Predicting Self-Reported Personality Traits

Regression	Personal Initiative				Conscientiousness			
	PI7	PI10	PI1	PI9	C15	C20	C18	C12
β s Trait Social Desirability (TSD)	.51*	.31	.61*	-.30	-.29	.60	-.14	.33
β s version i (without stems) x TSD	.08	.01	-.29	-.25	-.17	-.23	-.16	-.26
β s (beh. tendency) x TSD	-.07	-.33	-.28	.00	.43	.36	.03	-.05
β s (knowledge) x TSD	-.26	-.07	-.18	.25	.63*	.19	.46	.03
β s (desirability) x TSD	-.25	-.07	-.53	-.04	.65**	.28	.23	.15
β s version iii (open ended) x TSD	-.01	.00	.00	.00	-.01	.00	-.01	.00
β s version iv (full) x TSD	-.58*	-.09	.16	.35	-.15	-.37	.11	.14
Adj. R^2 s	.06	.03	.02	.10	.09	.05	.03	.05

Note. $N = 121$.

* $p < 0.05$. ** $p < 0.01$. *** $p < .001$.

RQ4: Prediction of Self- and Supervisor-Rated Job Performance

RQ4a. Hierarchical regressions examined how item versions i–iv predicted self- and supervisor-rated job performance (IRB and OCBI). Among self-ratings, explained variance for the full models ranged from negligible to moderate (e.g., for item PI9 predicting IRB: adj. $R^2 = .20$). For personal initiative, item versions ii (behavioral) emerged as the most relevant predictor for self-rated OCBI but showed a scattered pattern of positive and negative

relationships (β s between $-.41$ and $.34$, p s $< .05$). Item version iii was also a significant predictor in some items for both self-rated IRB and OCBI (e.g., for item PI9, $\beta = .04$, $p < .05$). For conscientiousness, item version ii (behavioral) was similarly relevant for both IRB and OCBI but also showed some unexpected negative relationships (e.g. for item C20: $\beta = -.39$, $p < .05$ for OCBI). With supervisor-rated job performance, version iii most often emerged as a significant predictor for personal initiative items, and version ii (social desirability) for conscientiousness items, but both only became significant in some items and overall explained variance remained low (adj. R^2 s between $.02$ and $.17$ for the full models).

RQ4b. Again, incremental validity of item versions was examined in both directions. When item versions i-iii were added to the full item predicting self-reports, we found incremental validity in only one item for IRB (e.g., item PI9, $\Delta R^2 = .22$, $p < .01$), with mostly item version ii (behavioral) being the significant predictor causing the change. For self-rated OCBI, we found incremental validity for two items (e.g., for item PI7, $\Delta R^2 = .12$, $p < .05$; $\beta_{ii\text{behavioral}} = -.41$, $p < .05$). When the full item was added to the base model containing versions i-iii predicting self-reports, it explained incremental validity in two items for OCBI (item C20: $\Delta R^2 = .06$, $p < .05$; $\beta_{\text{full Item}} = .38$, $p < .05$; item PI9: $\Delta R^2 = .06$, $p < .05$; $\beta_{\text{full Item}} = -.39$, $p < .05$), and in two items for IRB (item C15: $\Delta R^2 = .07$, $p < .05$; $\beta_{\text{full Item}} = .28$, $p < .05$; item C20: $\Delta R^2 = .10$, $p < .05$; $\beta_{\text{full Item}} = .47$, $p < .05$). For supervisor-rated job performance, item version iii added incremental value beyond the full item in some items (e.g., for item PI7 predicting IRB: $\Delta R^2 = .27$, $p < .05$; $\beta_{iii} = .32$, $p < .05$), as did item version ii (e.g., for item C20 predicting OCBI: $\Delta R^2 = .25$, $p < .05$; $\beta_{iidesirability} = -.63$, $p < .05$). The full item added incremental validity in only one item (for item PI10 predicting OCBI: $\Delta R^2 = .27$, $p < .05$; $\beta_{\text{full Item}} = .58$, $p < .05$).

Overall, results suggest that the predictive power of individual item versions for job performance varies considerably by item and rating source, with versions ii and iii showing

incremental value in some cases, while the full item rarely added incremental explanatory power, which contrasts with the results of our analyses of predictions of corresponding constructs.

H2: Incremental Contribution of Item Version iii Above and Beyond Version iv in Predicting Self- And Supervisor-Rated Job Performance

To test H2, we computed hierarchical regressions to examine whether version iii (open-ended) explained incremental variance in self- and supervisor-rated job performance beyond the full item (see Appendix D for all results). Significant incremental contributions of version iii were observed in only two personal initiative items. For self-rated outcomes, version iii accounted for incremental variance in item PI9 predicting IRB ($\Delta R^2 = .05, p < .05$) and OCBI ($\Delta R^2 = .06, p < .05$). For supervisor-rated outcomes, significant ΔR^2 values were observed in item PI7 predicting IRB ($\Delta R^2 = .16, p < .01$) and in item PI9 predicting OCBI ($\Delta R^2 = .45, p < .01$). No significant incremental effects of version iii were found for any conscientiousness items.

Overall, while we did find some incremental contribution of open-ended items over full items, these results do not support H2.

Discussion

General Discussion

In the present study, we strove to examine the relevance of different parts of SJTs and shed light onto each individual part's unique properties. For this, we chose a within-subject longitudinal design. Across four waves, participants responded to four different SJT item versions: version i (that is, just the response options without the situation description), version ii (that is, a randomized list of all SJT items' response options, presented like a questionnaire with three different instructions; behavioral, effectiveness, and social desirability), version iii (that is, only the situation description was provided with an open-

format response instruction), and version iv (that is, the full item with no manipulation). With this approach, we extended previous research by Kaminski et al. (2019), Leeds (2018), Krumm et al. (2015), and Schäpers, Mussel, et al. (2019) in testing different SJT permutations against each other to see which version performed best in terms of construct- and criterion-related validity, and which part of an SJT item showed the strongest relation to the full, unmanipulated item (thereby possibly identifying what part of an SJT actually drives SJT answering processes).

The Relevance of Response Options as a Set

Across eight different items, we found that item version i (just the response options with no situation description provided) proved to be the strongest predictor of the full, unmanipulated item in most cases, a pattern that also emerged in bivariate correlations on the scale level. This suggests that it is the response options in their entirety that primarily drive the response processes in a full SJT item. While this finding challenges traditional conceptualizations of SJTs as situational measures (Fan et al., 2016; Weekley et al., 2015), it falls right in line with more recent research currents as proposed by Schäpers, Lievens, et al. (2019) or Krumm et al. (2015) who theorized and empirically substantiated that SJT items are not as dependent on situation descriptions as formerly thought.

The Relevance of Individual Response Options

Following the works of Leeds (2018) and Kaminski et al. (2019), we also tested the capabilities of SJT response options on their own, i.e., in a presentation format similar to typical self-report questionnaires. Due to this similarity in format, we hypothesized a close relationship between item version ii and self-reported personality measures. While some item-level prediction of self-reported corresponding constructs did occur for version ii, these effects were not consistent across items, and H1 was therefore not supported. Additional, not preregistered correlation analyses confirmed that pattern on a scale level, suggesting that in

our data, no strong relationship exists between SJT item versions and self-reported personality measures. While some of this might be due to either item selection or sample size (see limitations section), it also fits into the now proverbial “hot mess”, i.e., the ongoing problem with SJTs’ construct validity (McDaniel et al., 2016). In contrast, version ii (particularly the behavioral and knowledge instructions) showed predictive utility for supervisor-rated job performance in a few items, which supports the idea that response options can contribute meaningfully when externally rated performance is the criterion. This aligns with findings by Freudenstein et al. (2020), who argued that response options may drive situation construal in SJTs. One potential explanation for the present pattern is that situation descriptions without a clear connection to the specific job context may introduce irrelevant variance or noise, whereas response options that focus on concrete behavioral tendencies may offer a more stable and generalizable link to performance-related constructs.

We did not find a pattern of significant interaction effects for trait social desirability, i.e., the individual disposition to conform to social norms and to present themselves positively (that is, according to these norms) to others (Fischer & Fick, 1993). In two instances, we found a significant interaction for trait social desirability and item version ii (once in the effectiveness and once in the social desirability instruction) for predicting self-report personality measures, but overall, only few significant interaction effects were found. Considering that historically, social desirability has been considered an important explanation for answering biases in personnel selection methods such as SJTs (Ones et al., 1996; Wiggins, 1966), these limited findings are somewhat unexpected. One technical explanation may be statistical power, as our sample was not large enough to reliably detect small interaction effects (see limitations section). However, our data may also suggest that TSD plays a more fundamental role in SJT responding that lies beneath of what interaction effects may or may not be able to measure: across multiple items, TSD showed strong main effects

for predicting self-reported personality measures. In some items (for example item PI7), the full item predicted self-reported personal initiative significantly less strongly when TSD was high. Also, even in the absence of significant interactions, intercorrelations between the different different version ii instructions were rather high, despite being designed to tap into different response logics. This convergence may indicate that test-takers' responding in these formats was guided less by the explicit instructions and more by an underlying drive to answer in a specific way, i.e., possibly a socially desirable way. Such a pattern aligns with the proposition by Kaminski et al. (2019), who argued that social desirability is not merely an external influence on SJT performance, but an integral component of how individuals process response options. Accordingly, our findings may reflect that TSD operates as an inherent response tendency in SJTs, affecting test-takers' behavior across conditions, and thus rarely emerges as a moderator because its influence is already built into the main response process.

The Relevance of Situation Descriptions With an Open Response Format

Based on the works of Rockstuhl et al. (2015), we hypothesized that open-format responses would be the best predictor for job performance measures. However, this hypothesis (H2) was not supported by our data. While item version iii showed incremental validity in a few isolated cases (mostly for personal initiative) it did not consistently predict job performance outcomes. Moreover, the overall explained variance remained low. This stands in contrast to prior findings on situational interviews, which are similar in format to open-ended SJTs and demonstrate strong criterion-related validity (Ingold et al., 2015; Jansen et al., 2013; Oostrom et al., 2016). A likely explanation lies in the nature of our item content: our SJT items were not chosen to specifically include clearly defined dilemma situations, which are thought to drive predictive value in open-response formats (Latham & Itzchakov, 2021; Latham & Sue-Chan, 1999).

Still, when considering all item versions, some results emerged that add to the body of research that finds general predictive properties of SJTs for job performance measures (Christian et al., 2010; McDaniel et al., 2007; McDaniel & Nguyen, 2001; Webster et al., 2020). Version ii (specifically the behavioral and knowledge instructions) was most frequently associated with self-rated job performance, albeit inconsistently. For supervisor-rated performance, version iii (open-ended) emerged most often as a significant predictor, particularly for personal initiative (but note that this analysis might have been underpowered). This contrast is notable, as bivariate correlations between self- and supervisor-rated job performance were substantial. One possible explanation is that self-ratings reflect internal trait representations, whereas supervisor ratings may tap into observable behaviors, thereby making open responses more diagnostic in the latter case (Lievens & Motowidlo, 2016). However, these findings should be interpreted with caution due to the comparatively small supervisor-rated subsample.

Implications for SJT Theory

While overall these results are less consistent than anticipated, they do fall into place in the context of the SJT response models proposed by Ployhart (2006), Grand (2020), and Martin-Raugh and Kell (2021). Ployhart's Response Process Model outlines four steps in responding to SJT items: comprehension, retrieval, judgment, and response selection. Our findings revealed that item version i (response options without situational context) explained a substantial amount of variance in the full SJT item, particularly for personal initiative items (see Table 4). Adding versions ii and iii only marginally increased explained variance in most cases, indicating that the response options together as intended may already contain sufficient situational information to support comprehension and judgment processes. In line with the reasoning of Fan et al. (2016), this could mean that test-takers are able to reconstruct situational meaning from the information presented within the response options, even in the

absence of situation descriptions. Within Ployhart's framework, this suggests that comprehension and retrieval may not require a fully fleshed-out situation in all cases, and that judgment and selection may be primarily driven by the content of the alternatives themselves.

Regarding the SiRJ framework, Grand's model posits that test-takers engage in conditional reasoning, similarity judgment, and preference accumulation when responding to SJT items. These processes are assumed to be activated by situational context and instructions that provide contextual framing. However, across item versions, our results offer only limited empirical support for these mechanisms. Version ii, which included only the response options in a randomized order and varied instruction types, yielded highly similar responses across the three instructions, suggesting that instruction framing had little influence on test-takers' response strategies (see Table 3). Version i, which provided some more situational context in that the response options were presented together (Melchers & Kleinmann, 2016), still predicted the full item version remarkably well. This pattern seems more consistent with heuristic response processes than with conditional reasoning. While version iii, which included an open-response format and full situational framing, did show incremental predictive value in selected items, these effects were not consistent across constructs or rating sources. Taken together, our findings suggest that the processes proposed in the SiRJ framework, if activated, appeared only limited and inconsistent within our study. It may be that the item designs, which lacked strong dilemma structures or clear motivational triggers, did not sufficiently engage the reasoning and judgment processes that the model assumes. However, it is also possible that cognitive reasoning models like SiRJ alone do not fully capture the dynamics at play in our data, specifically in relation to external criteria as test-takers were informed from the beginning that they would be asked to provide supervisor feedback. The Tripartite Model of SJT Responding (Martin-Raugh & Kell, 2021) includes motivational processes: it describes three sequential steps (i.e., situation perception, goal

formation, and response evaluation), which determine how individuals respond to SJT items. In our study, item version ii omitted the situation component. Across the three instruction types (behavioral, effectiveness, and social desirability), we observed highly similar response patterns (see Table 3), suggesting that response evaluation may have followed a relatively stable internal standard. At the same time, the behavioral instruction yielded stronger predictive validity than the other instructions, e.g., for self-rated job performance in several items. These findings suggest that, although instructions did not substantially shift surface-level response choice, they may have subtly shaped test-takers' internal goal orientation during the response process. In line with the Tripartite Model, such motivational processes could have influenced how well certain response styles aligned with different performance criteria. Thus, our findings point to a constrained but interpretable role of goal formation and evaluation processes in SJT responding, which may depend on the degree to which item format and instruction activate meaningful differentiation in response intent.

Practical Implications

Our findings open up some implications for the design and use of SJTs in personnel selection. First, the fact that response options alone (without situation descriptions) predicted full SJT items most strongly suggests that situational framing may not be essential for eliciting valid responses, supporting prior research by Krumm et al. (2015). Practitioners may therefore consider focusing more on the construction and quality of response options rather than assuming that elaborate situation descriptions are always necessary. Second, while open-format responses were not uniformly predictive within our study, they showed incremental validity for supervisor-rated job performance in selected cases. This implies that open-ended formats may be particularly useful when the criterion involves externally observable behavior rather than self-assessed traits.

Third, our results offer the possibility of social desirability not merely functioning as a biasing factor, but as an inherent part of SJT response processes, thereby affecting outcomes even in the absence of explicit instructions. Accordingly, test developers should treat trait social desirability as a construct to be understood and integrated, rather than automatically controlled for. Lastly, the high intercorrelation of responses across different instructions and the heterogeneity in item functioning suggest that careful consideration of instructional framing and improving item-level consistency should be key priorities in future SJT development.

Limitations

For the present study, it is important to acknowledge several limitations. First and foremost, we did not reach the a priori calculated sample sizes, specifically for moderation analyses and supervisor-related analyses; and post-hoc power analyses revealed less than sufficient power for these calculations. In part, this was to be expected due to the elaborate design of the study which required participants to not only follow through for four measurement times spanning two months, but also to have their supervisors fill out an additional questionnaire. In order to provide an alternative to supervisor ratings, we included self-rated job performance measures. We found substantial correlations between self- and supervisor ratings but acknowledge that this is a limitation to the study. Our moderation analyses were also underpowered. As stated above we still included the preregistered calculations, but advise interpreting the given results with caution, and to specifically not generalizing the findings for supervisor- and moderation analyses to other samples and/or SJTs.

In terms of generalizability, we strove to include items from different but equally established SJTs in order to cover a range of different SJTs, because of the fact that SJTs are notoriously heterogeneous even at the item level (Patterson et al., 2012; Sorrel et al., 2016).

By choosing the personal initiative and a personality SJT (Bledow & Frese, 2009; Mussel et al., 2018) we provided thoroughly pretested and valid construct-driven items (Freudenstein et al., 2020; Reznik et al., 2023; Schäpers et al., 2020). However, we were only able to include four items from each SJTs, as otherwise participant burden and processing time for each wave would have been considerably higher. We acknowledge that, while providing valuable knowledge about the different parts of SJTs, we cannot rely on this total of eight items to offer valid personality assessment, which is illustrated by the nonsignificant (or even significant but negative) relationships between SJT items and self-reported personality measures. Notably, we also included only construct-driven, personality-focused SJTs and did not incorporate any knowledge-focused SJTs. While we did include one item version with a knowledge instruction (and did find some substantial relationships between that version and job performance measures), we cannot make founded claims about knowledge- or effectiveness-focused SJTs, even though they are commonly found in personnel selection (Lievens & Patterson, 2011).

Future Research

The present study revealed a differentiated and partly inconsistent pattern of results, which points to several avenues for future research. Most notably, our design allowed for a theory-driven, within-subject comparison of multiple SJT item components. This revealed substantial item-level variability in predictive utility, suggesting that unified SJT response models may need to account for construct-specific and item-specific differences. We therefore call for extensive future research to further deepen our understanding on the individual properties of SJT item components. Specifically, it is important to broaden item selection as to include a wider variety of SJT domains in terms of constructs assessed. It would also be vital to gather a larger sample for further longitudinal studies in which not only different item versions should be presented, but that should also include real-world job

performance metrics in order to strengthen criterial and ecological validity. We also call for reconsideration of social desirability as both a bias and a commodity in SJTs. While we did not have sufficient power to assess a possible moderating relationship throughout, future studies should look into this again and also broaden the examination to different types of SJTs, as video-based or other forms of more interactive SJTs might bring forth an even more important role for social desirability than traditional text-based tests. Additionally, our findings suggest that social desirability may not merely influence responses in isolated conditions but may function as a stable underlying factor across formats and instructions, which highlights the need to integrate it more centrally into SJT theory and design.

Finally, we call for further refinement of open-format SJTs, specifically as predictors of job performance: in future studies, we should try and examine what processes are at work that make this format better at predicting supervisor-rated, but worse at predicting self-report job performance. Moreover, these analyses should ideally be conducted not only at the item level but extended to full SJT instruments in all version types, allowing researchers to evaluate version effects at the test level rather than within individual items.

Conclusion

This study examined the distinct contributions of SJT item components to construct and criterion validity. While the tested hypotheses were not supported, several interpretable patterns emerged. Response options alone often explained substantial variance in full items, suggesting that they play a central role in shaping SJT responses, possibly more so than situational stems. Predictive validity varied by criterion and item format, with behavioral instructions and open-ended responses showing incremental validity in selected cases.

Taken together, these findings underscore the heterogeneity of SJT functioning and highlight the need to account for item-level variance, response framing, and criterion type. Rather than supporting one uniformly superior format, our results point to a complex

interplay of item features. This complexity is mirrored in current process models of SJT responding (e.g., Ployhart, 2006; Grand, 2020; Martin-Raugh & Kell, 2021), which may require further refinement to fully capture the interaction between item design, motivational orientation, and response behavior.

References

- Bledow, R., & Frese, M. (2009). A 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>
- Bradley, K. M., & Hauenstein, N. M. A. (2006). The moderating effects of sample type as evidence of the effects of faking on personality scale correlations and factor structure. *Psychology Science, 48*(3), 313–335.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81.
<https://doi.org/https://doi.org/10.1037/h0046016>
- Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures: Interactionist psychology operationalized. In N. Christiansen & R. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). Routledge.
<https://doi.org/10.4324/9780203526910>
- Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A, Ford, C., Volcic, R., & De Rosario, H. (2020). *pwr* (Version 1.3-0) [R package]. GitHub.
<https://github.com/heliosdrm/pwr>
- 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>
- Fan, J., Stuhlman, M., Chen, L., & Weng, Q. (2016). Both general domain knowledge and situation assessment are needed to better understand how SJTs work [Commentary]. *Industrial and Organizational Psychology, 9*(1), 43–47.
<https://doi.org/10.1017/iop.2015.114>

- Fischer, D. G., & Fick, C. (1993). Measuring social desirability: Short forms of the Marlowe-Crowne social desirability scale. *Educational and Psychological Measurement*, 53(2), 417–424. <https://doi.org/10.1177/0013164493053002011>
- 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*, 73(4), 669–700. <https://doi.org/10.1111/peps.12385>
- Grand, J. A. (2020). A general response process theory for situational judgment tests. *Journal of Applied Psychology*, 105(8), 819–862. <https://doi.org/10.1037/apl0000468>
- Ingold, P. V., Kleinmann, M., König, C. J., Melchers, K. G., & Van Iddekinge, C. H. (2015). Why do situational interviews predict job performance? The role of interviewees' ability to identify criteria. *Journal of Business and Psychology*, 30(2), 387–398. <https://doi.org/10.1007/s10869-014-9368-3>
- 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), 1–27. <https://doi.org/10.1111/joop.12151>
- Jansen, A., König, C. J., Stadelmann, E. H., & Kleinmann, M. (2012). Applicants' self-presentational behavior: What do recruiters expect and what do they get? *Journal of Personnel Psychology*, 11(2), 77–85. <https://doi.org/10.1027/1866-5888/a000046>
- Jansen, A., Melchers, K. G., Kleinmann, M., Lievens, F., 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>

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), 399–416. <https://doi.org/10.1037/a0037674>

Latham, G. P., & Itzhakov, G. (2021). The effect of a dilemma on the relationship between ability to identify the criterion (ATIC) and scores on a validated situational interview.

Frontiers in Psychology, 12. <https://doi.org/10.3389/fpsyg.2021.674815>

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>

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>

Leeds, J. P. (2018). Applying cognitive acuity theory to the development and scoring of situational judgment tests. *Behavior Research Methods*, 50(6), 2215–2225.

<https://doi.org/10.3758/s13428-017-0988-1>

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and*

Organizational Psychology, 9(1), 3–22. <https://doi.org/10.1017/iop.2015.71>

Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance

- in advanced-level high-stakes selection. *Journal of Applied Psychology*, 96(5), 927–940. <https://doi.org/10.1037/a0023496>
- Martin-Raugh, M. P., & Kell, H. J. (2021). A process model of situational judgment test responding. *Human Resource Management Review*, 31(2), 100731. <https://doi.org/10.1016/j.hrmr.2019.100731>
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60(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 [Commentary]. *Industrial and Organizational Psychology*, 9(1), 47–51. <https://doi.org/10.1017/iop.2015.115>
- 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>
- McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology*, 36(4), 979–1016. <https://doi.org/10.1111/j.0021-9029.2006.00052.x>
- Melchers, K. G., & Kleinmann, M. (2016). Why situational judgment is a missing component in the theory of SJTs [Commentary]. *Industrial and Organizational Psychology*, 9(1), 29–34. <https://doi.org/10.1017/iop.2015.111>
- 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>
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences, 106*, 183–189. <https://doi.org/10.1016/j.paid.2016.11.014>
- Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*(6), 660. <https://doi.org/10.1037/0021-9010.81.6.660>
- Oostrom, J. K., Melchers, K. G., Ingold, P. V., & Kleinmann, M. (2016). Why do situational interviews predict performance? Is it saying how you would behave or knowing how you should behave? *Journal of Business and Psychology, 31*, 279–291. <https://doi.org/10.1007/s10869-015-9410-0>
- Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education, 46*(9), 850–868. <https://doi.org/10.1111/j.1365-2923.2012.04336.x>
- 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.
- 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., & Sherman, R. A. (2015). Measuring the situational eight DIAMONDS characteristics of situations. *European Journal of Psychological Assessment, 32*(2), 155–164. <https://doi.org/10.1027/1015-5759/a000246>
- Revelle, W. (2025). *psych* (Version 2.5.3) [R package]. CRAN. <https://cran.r-project.org/package=psych>
- Reznik, N., Krumm, S., Freudenstein, J. P., Heimann, A. L., Ingold, P., Schäpers, P., & Kleinmann, M. (2023). Does understanding what a test measures make a difference? On the relevance of the ability to identify criteria for situational judgment test performance. *International Journal of Selection and Assessment, 32*(2), 210–224. <https://doi.org/10.1111/ijsa.12458>
- 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>
- Rosen, N. A. (1961). How supervise?—1943–1960. *Personnel Psychology, 14*(1), 87–99. <https://doi.org/10.1111/j.1744-6570.1961.tb00925.x>
- Salgado, J. F., & Táuriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3–30. <https://doi.org/10.1080/1359432X.2012.716198>
- Schäpers, P., Freudenstein, J.-P., Mussel, P., Lievens, F., & Krumm, S. (2019). Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests. *Journal of Research in Personality, 87*, 103963. <https://doi.org/10.1016/j.jrp.2020.103963>

- Schäpers, P., Krumm, S., Lievens, F., Freudenstein, J.-P., Schulze, J., & König, C. J. (2022, Jan 11-14). *Situation Descriptions in Situational Judgment Tests: A Matter of Trait Activation?* [Conference presentation]. European Association of Work and Organizational Psychology, Glasgow, Scotland.
- 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*, 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 perceptions. *Journal of Applied Psychology*, 105(8), 800–818. <https://doi.org/10.1037/apl0000457>
- Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, 19(3), 506–532. <https://doi.org/10.1177/1094428116630065>
- Webster, E. S., Paton, L. W., Crampton, P. E., & Tiffin, P. A. (2020). Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education*, 54(10), 888–902. <https://doi.org/10.1111/medu.14201>
- 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>
- Wiggins, J. S. (1966). Social desirability estimation and "faking good" well. *Educational and Psychological Measurement*, 26(2), 329–341. <https://doi.org/10.1177/001316446602600206>

- 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(3), 601–617. <https://doi.org/10.1177/014920639101700305>
- Ziegler, M. (2011). Applicant faking: A look into the black box. *The Industrial and Organizational Psychologist*, 49, 29-36.
- Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods*, 18(4), 679–703. <https://doi.org/10.1177/1094428115574518>
- Ziegler, M., MacCann, C., & Roberts, R. D. (Eds.). (2012). *New perspectives on faking in personality assessments*. Oxford University Press.

Appendix A

Item Versions

Item Version i (Without Description)

- Situation description: /
- Instruction: What would you do?
- Response options:
- a) Behavior 1
 - b) Behavior 2
 - c) Behavior 3
 - d) Behavior 4

Item Version ii (Randomized List of Response Options)

- Situation description: /
- Instruction: I would behave like this. (behavioral)
This behavior is effective. (knowledge)
This behavior is socially desirable. (social desirability).
- Response options:
- a) Behavior 1 (Item C12)
 - b) Behavior 4 (Item PI1)
 - c) Behavior 2 (Item PI9)
 - d) Behavior 2 (Item C15)
 - e) Behavior 3 (Item C20)
 - f) Behavior 3 (Item PI7)
 - g) Behavior 1 (Item PI10)
 - h) Behavior 2 (Item C18)

...

Item Version iii (Open-Ended)

Situation description: Description of a realistic situation

Instruction: What would you do?
(open response field)

Response options: /

Item Version iv (Full Item)

Situation description: Description of a realistic situation

Instruction: What would you do?

Response options: a) Behavior 1
b) Behavior 2
c) Behavior 3
d) Behavior 4

Appendix B

T-Tests

Table B1

Welch's t-Test Results for Sequence Effects Between Groups With Different Wave Sequence

Test	$t(df)$	Mean difference	95% CI for mean difference	p -value b	p
A vs. A1, C-i	$t(332.13) = 0.44$	0.76	[-2.62, 4.14]	1.0	.659
A vs. A1, PI-i	$t(330.35) = 1.00$	2.37	[-2.28, 7.01]	1.0	.317
B vs. B1, PI-iib	$t(155.03) = 2.42$	0.13	[0.02, 0.24]	.14	.017*
B vs. B1, PI-iie	$t(140.68) = 1.64$	0.10	[-0.02, 0.22]	.82	.103
B vs. B1, PI-iisd	$t(142.74) = 1.03$	0.06	[-0.05, 0.16]	1.0	.307
B vs. B1, C-iib	$t(130.08) = 0.50$	0.04	[-0.11, 0.18]	1.0	.620
B vs. B1, C-iie	$t(128.13) = -0.02$	-0.001	[-0.15, 0.15]	1.0	.988
B vs. B1, C-iisd	$t(139.22) = -0.11$	-0.01	[-0.14, 0.12]	1.0	.912

Note. Significant tests are marked with an asterisk (*). P -values in p -value b are Bonferroni-corrected for multiple testing. A, A1, B, B1 = participant groups with different wave sequences. PI = Personal Initiative SJT. C = Conscientiousness SJT. i = item version i. ii = item version ii. b = behavioral instruction, e = effectiveness instruction, sd = social desirability instruction.

Table B2

Welch's t-Test Results for Learning Effects Between Groups With and Without Previous Wave Exposure

Test	$t(df)$	Mean difference	95% CI for mean difference	p -value b	p
Cad vs Cr, PI-iii	$t(77.86) = 1.09$	0.11	[-0.09, 0.31]	1.0	.278
Cad vs Cr, C-iii	$t(76.80) = 0.73$	0.07	[-0.12, 0.27]	1.0	.469
Dad vs Dr, PI-iv	$t(121.47) = 0.57$	0.05	[-0.12, 0.21]	1.0	.569
Dad vs Dr, C-iv	$t(89.29) = 1.58$	0.16	[-0.04, 0.35]	.47	.118

Note. Significant tests are marked with an asterisk (*). P -values in p -value b column are Bonferroni-corrected for multiple testing. Cad = additional group for wave C, Cr = regular group for wave C, Dad = additional group for wave D, Dr = regular group for wave D. PI = Personal Initiative SJT, C = Conscientiousness SJT. iii = item version iii, iv = item version iv.

Appendix C

Relative Weights for SJT Item Versions i to iii in Predicting the Full SJT Item Version

Item versions	Personal Initiative SJT				Personality SJT			
	(version iv; full item)				(version iv; full item)			
	7	10	1	9	15	20	18	12
Version i (without stems)	68.37	50.29	76.61	63.45	9.38	63.99	68.01	24.16
Version ii (beh. tendency)	23.35	27.99	13.86	15.34	41.37	27.06	8.73	19.44
Version ii (knowledge)	6.14	17.83	2.56	3.31	27.26	5.86	14.59	51.65
Version ii (desirability)	1.91	0.76	5.34	16.08	11.28	2.72	2.44	4.09
Version iii (open ended)	0.21	3.12	1.62	1.82	10.71	0.38	6.23	0.66

Note. Numbers are percentages.

Appendix D

Regressions

Table D1a

Hierarchical Regression Analyses with Personal Initiative SJT Item Versions i to iv Predicting Self- and Supervisor Reports

Regression Steps	Self-report			Supervisor report	
	Personal Initiative	IRB	OCBI	IRB	OCBI
	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9
Step 1: Version iv (full)					
βs	-.02 / .29* / .05 / -.66***	.29* / .15 / .21 / -.26	.0002 / .27 / .13 / -.3*	-.009 / -.07 / -.12 / .2	-.2 / .45* / -.13 / -.2
Adj. R ² s	.009 / .05 / .02 / .17	.04 / .005 / .02 / .03	.01 / .03 / .004 / .04	.02 / .04 / .01 / .01	.02 / .08 / .0001 / .01
Step 2: Version i, versions ii, and version iii					
βs version i (without stems)	.02 / -.32 / .29 / .21	-.11 / .37 / -.09 / -.14	.05 / -.15 / -.07 / .05	.09 / .33 / .04 / .14	.16 / -.25 / -.13 / .34
βs version ii (beh. tendency)	-.27 / -.19 / .44* / .05	.006 / .03 / .28 / .14	-.41* / -.15 / .23 / .34*	-.04 / .09 / .05 / -.04	-.5* / .1 / .04 / -.14
βs version ii (knowledge)	.03 / .15 / -.44* / .02	.11 / -.11 / .2 / -.49***	-.02 / .03 / .23 / -.48**	.25 / -.33 / .21 / .18	.11 / -.5 / -.27 / .41
βs version ii (desirability)	.24 / -.22 / .01 / -.32	.26 / .15 / .26 / .07	.28 / .17 / .21 / -.12	.04 / -.06 / .05 / .09	-.08 / .02 / -.12 / -.08
βs version iii (open ended)	-.006 / .06 / .02 / .03	.03 / .08 / -.004 / .04*	.04 / .37 / .02 / .04*	.32* / .42 / -.24 / .08**	-.11 / .76 / .02 / .07**
Adj. R ² s	.007 / .03 / .04 / .19	.08 / .01 / .04 / .2	.07 / .003 / .01 / .2	.14 / .05 / .15 / .52	.08 / .2 / .02 / .43
ΔR ²	.05 / .11 / .15 / .08	.08 / .12 / .11 / .22**	.12* / .10 / .1 / .21**	.27* / .31 / .11 / .59**	.16 / .31 / .11 / .53**
ΔF	1.06 / 3.67 / 1.68 / 1.26	1.8 / .86 / 1.29 / 3.8	2.64 / .78 / 1.17 / 3.6	2.55 / 1.4 / .45 / 6.48	1.5 / 1.6 / .48 / 4.86

Chapter 3: SJTs and Their Components

Note. Post-hoc power analyses revealed that, due to less than expected return on supervisor acquisition, these regressions have less than sufficient power. With $N = 61$ complete cases, 4 predictors, and a significance level of $\alpha = .05$, the analysis had a power of 0.65 to detect a medium effect ($f^2 = 0.23$). As stated above, we advise to interpret findings in this section with caution. $N_{\text{self}} = 121$, $N_{\text{supervisor}} = 61$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table D1b

Hierarchical Regression Analyses with Personal Initiative SJT Item Versions i to iv (Reversed Steps) Predicting Self- and Supervisor Report

Regression Steps	Self-report					Supervisor report				
	Personal Initiative		IRB		OCBI		IRB		OCBI	
	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	PI7 / PI10 / PI1 / PI9	
Step 1: Version i, versions ii, and version iii										
βs version i (without stems)	.04 / -.18 / .23 / -.15	.02 / .39* / -.03 / -.24	.01 / -.02 / -.04 / -.15	.11 / .34 / .01 / .29	.02 / .04 / -.15 / .30					
βs version ii (beh. tendency)	-.27 / -.16 / .40 / .07	.06 / .04 / .33 / .14	-.39* / -.13 / .25 / .35*	-.05 / .08 / -.01 / -.03	-.50* / -.30 / .00 / -.14					
βs version ii (knowledge)	-.03 / .10 / -.41 / .07	.13 / -.12 / .16 / -.48**	.02 / -.02 / .21 / -.46**	.25 / -.32 / .30 / .03	.11 / -.33 / -.23 / .46					
βs version ii (desirability)	.24 / -.22 / .09 / -.46*	.28 / .15 / .26 / .03	.29 / .18 / .21 / -.21	.05 / -.07 / -.01 / .17	.03 / -.21 / -.16 / -.11					
βs version iii (open ended)	-.01 / .02 / .02 / .02	.03 / .08 / .00 / .04*	.04 / .34 / .03 / .04	.31* / .40 / -.24 / .08**	-.06 / .26 / .02 / .07**					
Adj. R ² s	.00 / -.09 / .05 / .05	.06 / .01 / .04 / .19	.07 / -.08 / .03 / .14	.17 / .11 / -.15 / .49	.08 / -.08 / -.08 / .45					
Step 2: Version iv (full)										
βs version i (without stems)	.02 / -.32 / .30 / .21	-.11 / .37 / -.09 / -.14	.05 / -.15 / -.08 / .05	.10 / .33 / .04 / .14	.16 / -.25 / -.13 / .35					
βs version ii (beh. tendency)	-.28 / -.19 / .44 / .05	.01 / .04 / .29 / .14	-.41* / -.15 / -.23 / .34*	-.05 / .09 / .06 / -.03	-.50* / .10 / .04 / -.14					
βs version ii (knowledge)	-.03 / .15 / -.45 / .02	.11 / -.11 / .20 / -.49**	.02 / .03 / -.23 / -.48**	.25 / -.33 / .21 / .19	.11 / -.50 / -.27 / .42					
βs version ii (desirability)	.24 / -.22 / .10 / -.32	.26 / .15 / .26 / .07	.28 / .17 / .21 / -.13	.04 / -.06 / .05 / .09	.08 / .02 / -.12 / -.08					
βs version iii (open ended)	-.01 / .06 / .02 / .03	.03 / .09 / .00 / .04*	.04 / .38 / .03 / .04	.32* / .42 / -.24 / .08**	-.12 / .76 / .02 / .07**					
βs version iv (full)	.04 / .39* / -.16 / -.69**	.27 / .06 / .15 / -.21	.09 / .37 / .09 / -.39*	.02 / .04 / -.18 / .30	-.19 / .58* / -.09 / -.09					
Adj. R ² s	-.01 / .04 / .04 / .19	.08 / -.02 / .04 / .20	.07 / .00 / .01 / .20	.14 / .05 / -.15 / .52	.08 / .20 / -.13 / .43					
ΔR ²	.00 / .13* / .01 / .14**	.03 / .00 / .01 / .02	.00 / .09 / .01 / .06*	.00 / .00 / .00 / .04	.00 / .27* / .01 / .00					
ΔF	.07 / 4.93 / .79 / 11.39	3.36 / .15 / .08 / 1.56	.30 / 3.51 / .28 / 5.0	.02 / .03 / .85 / 1.99	.78 / 7.15 / .31 / .15					

Chapter 3: SJTs and Their Components

Note. Post-hoc power analyses revealed that, due to less than expected return on supervisor acquisition, these regressions have less than sufficient power. With $N = 61$ complete cases, 4 predictors, and a significance level of $\alpha = .05$, the analysis had a power of 0.65 to detect a medium effect ($f^2 = 0.23$). As stated above, we advise to interpret findings in this section with caution. $N_{\text{self}} = 121$, $N_{\text{supervisor}} = 61$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table D2a

Hierarchical Regression Analyses with Conscientiousness SJT Item Versions i to vi Predicting Self- and Supervisor Reports

Regression Steps	Self-report			Supervisor report	
	Conscientiousness	IRB	OCBI	IRB	OCBI
	C15 / C20 / C18 / C12	C15 / C20 / C18 / C12	C15 / C20 / C18 / C12	C15 / C20 / C18 / C12	C15 / C20 / C18 / C12
Step 1: Version iv (full)					
βs	.11 / .27 / .2 / .02	.27** / .34* / .25 / .18	-.13 / .29* / .03 / .005	.16 / .05 / .26 / -.02	-.08 / .31 / .33 / .03
Adj. R ² s	.002 / .02 / .01 / .00	.07 / .05 / .02 / .01	.00 / .03 / .01 / .001	.02 / .02 / .01 / .02	.02 / .01 / .01 / .02
Step 2: Version i, versions ii, and version iii					
βs version i (without stems)	-.27 / -.32* / .04 / -.14	-.12 / -.20 / .14 / -.19	-.13 / -.004 / -.13 / .12	-.55** / -.20 / .43* / -.08	-.68* / -.15 / .4 / -.24
βs version ii (beh. tendency)	-.34* / -.20 / .04 / -.01	.16 / -.21 / .15 / .15	.06 / -.39** / .04 / -.17	.11 / -.16 / -.2 / .26	.07 / -.39 / -.2 / -.28
βs version ii (knowledge)	.25 / .23 / .11 / .31*	-.08 / .13 / .03 / .14	-.04 / .14 / -.04 / -.05	.11 / -.38* / .17 / .18	-.19 / .62* / .27 / .25
βs version ii (desirability)	.08 / .14 / .01 / -.23	-.13 / .07 / -.19 / -.22	-.22 / .16 / -.27 / -.11	-.12 / -.27 / -.29 / -.43*	-.03 / -.10 / -.22 / -.63*
βs version iii (open ended)	-.01 / .00 / -.02 / -.00	-.01 / -.02 / -.01 / -.003	-.03 / -.02 / -.03 / .003	-.02 / -.16 / -.21 / .01	-.02 / .42 / -.24 / .00
Adj. R ² s	.05 / .06 / .002 / .01	.06 / .07 / .00 / .01	.02 / .08 / .01 / .02	.17 / .07 / .05 / .04	.11 / .17 / .02 / .02
ΔR ²	.09 / .08 / .02 / .06	.04 / .06 / .00 / .05	.06 / .09 / .06 / .05	.24* / .19 / .15 / .18	.22 / .25* / .08 / .17
ΔF	2.07 / 1.69 / .37 / 1.02	.79 / 1.30 / .61 / .87	1.24 / 2.00 / 1.3 / .75	2.67 / 1.87 / 1.42 / 1.48	2.24 / 2.73 / .72 / 1.37

Note. Post-hoc power analyses revealed that, due to less than expected return on supervisor acquisition, these regressions have less than sufficient power. With $N = 61$ complete cases, 4 predictors, and a significance level of $\alpha = .05$, the analysis had a power of 0.65 to detect a medium effect ($f^2 = 0.23$). As stated above, we advise to interpret findings in this section with caution. $N_{\text{self}} = 121$, $N_{\text{supervisor}} = 61$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table D2b

Hierarchical Regression Analyses with Personal Initiative SJT Item Versions i to iv (Reversed Steps) Predicting Self- and Supervisor Reports

Regression Steps	Self-report			Supervisor report	
	Conscientiousness	IRB	OCBI	IRB	OCBI
	C15 / C20 / C18 / C12	C15 / C20 / C18 / C12	C15 / C20 / C18 / C12	C15 / C20 / C18 / C12	C15 / C20 / C18 / C12
Step 1: Version i, versions ii, and version iii					
βs version i (without stems)	-.27 / -.20 / .08 / -.16	-.13 / -.06 / .20 / -.17	-.13 / .11 / -.11 / .13	-.49* / -.14 / .47* / -.07	-.68* / -.05 / .45 / -.23
βs version ii (beh. tendency)	-.29 / -.13 / .05 / -.02	.26 / -.14 / .16 / .14	.02 / -.33* / .04 / -.17	.19 / -.13 / -.21 / .24	.11 / -.35 / -.20 / .27
βs version ii (knowledge)	.30 / .21 / .13 / .27	-.01 / .11 / .05 / .20	-.07 / .13 / -.03 / -.02	.20 / .37* / .20 / .14	-.15 / .61* / .31 / .22
βs version ii (desirability)	.09 / .15 / .01 / -.21	-.13 / .08 / -.20 / -.25	-.22 / .17 / -.27 / -.12	-.18 / -.28 / -.27 / -.42*	-.05 / -.12 / -.20 / -.63*
βs version iii (open ended)	-.02 / .00 / -.02 / .00	-.02 / -.02 / -.01 / .00	-.03 / -.02 / -.03 / .00	-.02 / -.08 / -.21 / .01	-.02 / .56 / -.24 / .00
Adj. R ² s	.04 / .00 / -.03 / .00	.00 / -.02 / .00 / .01	.02 / .03 / .01 / -.02	.14 / .07 / .07 / .04	.12 / .17 / .00 / .05
Step 2: Version iv (full)					
βs version i (without stems)	-.27 / -.32* / .04 / -.14	-.12 / -.20 / .14 / -.20	-.13 / .00 / -.13 / .12	-.55* / -.20 / .43 / -.08	-.70* / -.15 / .40 / -.24
βs version ii (beh. tendency)	-.35* / -.20 / .04 / -.02	.16 / -.21 / .15 / .15	.06 / -.39** / .04 / -.17	.11 / -.16 / -.20 / .26	.08 / -.39 / -.20 / .28
βs version ii (knowledge)	.24 / .23 / .11 / .31	-.08 / .13 / .03 / .14	-.04 / .14 / -.04 / -.05	.11 / .38* / .17 / .18	-.19 / .62* / .27 / .25
βs version ii (desirability)	.09 / .14 / .01 / -.22	-.13 / .07 / -.19 / -.22	-.22 / .16 / -.27 / -.11	-.12 / -.27 / -.29 / -.43*	-.03 / -.10 / -.22 / -.63*
βs version iii (open ended)	-.01 / .00 / -.02 / .00	-.01 / -.02 / -.01 / .00	-.03 / -.02 / -.03 / .00	-.02 / -.16 / -.21 / .01	-.02 / .42 / -.24 / .00
βs version iv (full)	.16 / .41* / .13 / -.08	.28 / .47** / .18 / .13	-.10 / .38** / .08 / .06	.20 / .19 / .13 / -.12	.09 / .31 / .16 / -.08
Adj. R ² s	.05 / .06 / -.02 / .00	.06 / .06 / .00 / .01	.02 / .08 / .01 / -.03	.17 / .07 / .05 / .04	.11 / .17 / -.03 / .02
ΔR ²	.02 / .06* / .00 / .00	.07** / .10** / .00 / .00	.00 / .06* / .00 / .00	.04 / .18 / .01 / .02	.00 / .02 / .00 / .00
ΔF	2.28 / 6.63 / .55 / .26	7.83 / 10.47 / 1.22 / .85	.72 / 6.71 / .23 / .14	2.25 / .89 / .28 / .69	.21 / 1.24 / .21 / .14

Chapter 3: SJTs and Their Components

Note. Post-hoc power analyses revealed that, due to less than expected return on supervisor acquisition, these regressions have less than sufficient power. With $N = 61$ complete cases, 4 predictors, and a significance level of $\alpha = .05$, the analysis had a power of 0.65 to detect a medium effect ($f^2 = 0.23$). As stated above, we advise to interpret findings in this section with caution. $N_{\text{self}} = 121$, $N_{\text{supervisor}} = 61$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Chapter 4

SJTs and Their Situation Construal

It's a situation! It's a simulation! No - It's an SJT! On the Perceived Situation

Characteristics in Deconstructed Situational Judgement Test Items

Nomi Reznik¹, Stefan Krumm¹, Alyce Thiel¹, Jan-Philipp Freudenstein²

¹Department of Education and Psychology, Freie Universität Berlin

² Hogrefe Verlagsgruppe GmbH, Group R & D, Germany

This research was supported by a grant (KR 3457/2-2) from the German Research Foundation (DFG).

Abstract

Situational Judgment Tests (SJTs) are a staple of personnel selection, intended to simulate work-relevant situations and elicit trait-driven behavioral responses. Despite their widespread use, the extent to which SJTs function as true situational simulations (i.e., engaging test-takers in psychologically meaningful construal) remains yet unresolved. Grounded in the Situation Construal Model and Trait Activation Theory, the present preregistered study investigates whether test-takers' construal of situational characteristics predicts SJT item performance, and whether this relationship depends on the presence of trait-relevant cues. Test-takers ($N = 2,341$) were randomly assigned to one of seven experimental conditions that varied the presence of contextual information, trait-relevant cues, and response options in six SJT items. Participants rated each item on dimensions from the DIAMONDS taxonomy, situational strength, and perceived social desirability. Results from item-level mixed-effects regression analyses revealed that situation construal variables explained only a small portion of variance in SJT performance, with inconsistent predictive patterns across items. Interaction effects with trait-relevant cues were limited and non-systematic. Contrary to theoretical expectations, social desirability had minimal influence. These findings call into question key assumptions of SJT response process models, particularly the presumed centrality of situation construal. The results suggest that many SJTs may lack sufficient fidelity to elicit psychologically rich situational engagement, functioning more as contextualized knowledge assessments than as behavioral simulations. Implications for test development, including the design and placement of situational information, are discussed alongside theoretical considerations for the role of situational factors in judgment and decision-making.

Keywords: Situation Construal, Situational Judgment Test, Trait Activation Theory

Introduction

Is this situation the same for me as it is for you? The question on how individuals actually perceive situations, and how their behavior is shaped by their perceptions, has been asked again and again in the past millennium. Research in personality psychology has increasingly emphasized not only the influence of traits but also the dynamic and situationally dependent nature of human behavior (Fleeson, 2001; Funder, 2006; Mischel, 1968). The Situation Construal Model (Funder, 2016) posits that behavior is shaped by the interplay of personality traits, objective situational cues, and the individual's subjective interpretation of those cues, which Funder describes as situation construal. This model is further supported by frameworks such as the DIAMONDS taxonomy (Rauthmann et al., 2014), and provides a structured account of psychologically meaningful pieces of situational information, and distinctions between objective situational features ("cues"), their psychological interpretation ("characteristics"), and broader categories of situations ("classes"; Rauthmann et al., 2015; Reis, 2008).

These conceptual developments have influenced applied psychological domains, including simulation-based personnel selection procedures. Methods such as assessment centers and situational interviews use situational information to evoke job-relevant behavior, often relying on situations to constrain variability in response behavior and enhance interpretability (Lievens, 2017; Meyer & Dalal, 2009). The understandability and consistency of these situations (that is, their situational strength) can significantly shape how individuals perceive it and whether and what traits are expressed (Meyer et al., 2010). Trait Activation Theory (Tett & Burnett, 2003) further emphasizes that the activation of trait-relevant behavior depends on the presence of cues, which are theorized to serve as catalysts for trait expression.

Building on these principles, situational judgment tests (SJTs) are designed to present standardized, job-relevant scenarios including behavioral responses (Motowidlo et al., 1990). Although modeled after real-life situations, the extent to which SJTs function as true situational simulations remains a topic of empirical investigation (Krumm et al., 2015a). Recent research suggests that situation construal processes, including the identification of situational characteristics, trait-relevant cues, and perceptions of situational strength, may influence how test-takers navigate SJT items (Brown et al., 2016; Schäpers et al., 2020; Serfass & Sherman, 2013). However, empirical findings on these effects have been mixed, raising questions about whether SJTs effectively elicit the situational perception they are assumed to measure (Freudenstein, Schäpers, et al., 2023; Freudenstein et al., 2020; Jackson et al., 2017).

The present study contributes to this growing literature by systematically examining whether and how different conceptualizations of situation construal (that is, DIAMONDS, situational strength, and social desirability perception) predict performance on SJT items. Further, we assess whether the predictive value of these construal dimensions depends on the presence of trait-relevant cues. In doing so, we aim to clarify the mechanisms underlying situational judgment test performance and to explore whether SJTs truly function as simulations of psychologically meaningful situations.

Theoretical Background

From person and situation to person-situation interaction models

The idea of a perfectly dichotomous division of person and situation has been abandoned for the better part of the past five decades, making way for a more interactionist view on situation perception (Mischel, 1968). Funder (2006) specifically, some 20 years ago, proposed for research to conceptualize individuals in situations as a triad of the person, the situation, and their interaction, which inspired an entire field of person-situation-interaction

research. Several points are key when addressing person-situation-interaction: Making changes to a specific situation will result in possibly great changes in individuals' behavior (Funder & Colvin, 1991), from which we gather that situational context does drive and influence behavior somehow. Still, there is some consistency in individuals across situations; for example averaged trait levels remain stable throughout different contexts even though specific behaviors change (Fleeson, 2001). Thus, it is understood that both intraindividual consistency and variability in behavior can be found between situations; consistency being commonly linked to broader, more stable traits such as the Big Five (McCrae & Costa, 1987) and to more spontaneous, uncontrolled behaviors such as laughing or crying, and variability being linked to more specific, controlled and conscious behaviors such as starting an argument (Mischel & Shoda, 1995).

The Situation Construal Model

Building on the interactionist foundation, the Situation Construal Model provides a systematic framework to understand how situations are perceived and interpreted. In order to really ascertain what specific properties of a situation elicit some behaviors but not others, and what kinds of situations a trait can be stable across, Funder (2016) proposed the Situation Construal Model: a comprehensive framework where behavior is thought of as a function of personality traits, objective properties of a situation, and the way individuals construe that situation around them. Following the logic of Rauthmann et al. (2015), the specific properties or determinants of situations can be understood by identifying their *cues*, *characteristics*, and *classes*. *Cues* are defined as objective features of a given situation, e.g., if another person is present, the time of day, and the type of location the situation is set in. These are objectively measurable bits of data in a given environment that form the first building block of interpretable situation information (Reis, 2008): Individuals perceiving the situation will likely identify these cues all the same, but possibly vary in their interpretation (Rauthmann et

al., 2015). This interpretation of the objective cues is called *characteristics*: they constitute the psychologically meaningful attribution of individuals' personal knowledge, values, or other information (Brown et al., 2016) to the specific cue. With the help of situation characteristics, individuals are able to make the cognitive step from e.g., identifying a situation with another person in it, to identifying a situation with another person in it with whom they want to be romantically involved. Finally, *classes*, or higher-order categories of situations, group together situations based on shared cues or characteristics, e.g. situations at work or situations with a romantically interesting person in it. Mental representation of classes enable individuals to compare situations across different contexts (Rauthmann et al., 2015).

An extensive framework for situation characteristics was put forward by Rauthmann et al. (2014), who identified the Situational Eight DIAMONDS, that is, eight dimensions of situational features: Duty (obligation to perform tasks), Intellect (opportunities for intellectual engagement), Adversity (negative aspects like threat or criticism), Mating (opportunities for romantic interaction), pOsitivity (enjoyment of the situation), Negativity (potential for anxiety or stress), Deception (possibility of manipulation or deceit), and Sociality (social interaction and relationship development). The DIAMONDS model was found to effectively predict behavior in situations and provides a thorough explanation of individuals' situation perception (Rauthmann & Sherman, 2015).

Situational Strength

Another important framework for understanding situational influence is the concept of situational strength (Meyer et al., 2010), which is also directly related to behavior in situations: Situational strength refers to the degree to which external cues or pressures in a situation can be understood the same across individuals. In high-strength situations, behavior tends to be more uniform as these situations provide clear guidance on how to act, while low-

strength situations, characterized by ambiguity, allow for more variability in behavior across individuals (Meyer & Dalal, 2009; Meyer et al., 2010). Meyer and Dalal (2009) define situational strength as external situational "pressure" that constrains the influence of individual traits on behavior, following the framework of Mischel and Shoda (1995). Situational strength is typically understood as containing four key dimensions: *constraints*, which limit decision-making freedom; *consequences*, which refer to the importance of actions; *clarity*, or the extent to which expected behaviors are clearly communicated; and *consistency*, indicating how reliably behaviors are reinforced (Hough & Oswald, 2008; Meyer et al., 2014). For the present study, we understand situational strength as the extent to which interpretation of an objective situational cue is possible and necessary for the individual, e.g. a red traffic light poses a strong situational cue where almost all individuals will come to the same interpretation, and thus, behavior (that is, stopping). At the same time, a yellow traffic light would pose a weaker situational cue, where individuals would then need to take other personal knowledge, information, and values into account to decide whether they would stop or not (Schäpers et al., 2020). Notably, research by Ingold et al. (2015) showed that in structured situational interviews, the presence of clear situational cues contributes to a higher level of situational strength. In such contexts, individuals who possess the ability to identify, interpret, and name the demands the situation imposes on them (referred to as "ability to identify criteria" or "ATIC"; Kleinmann, 1993) tend to perform better.

Social Desirability Perception

Another situational determinant is the perception of social desirability: how much a response is perceived as socially favorable, which can influence choices even in ambiguous contexts (Kaminski et al., 2019). Leeds (2018) conceptualized that social desirability comprises a process of situation construal in which test-takers interpret what behavior is

socially appropriate in the given situation, thereby functioning similarly to situational strength by influencing individual behavior based on shared social norms instead of individual traits: When social desirability perception is high, most people behave in socially appropriate ways regardless of their individual trait levels, just as strong situations suppress individual trait expression (Fischer & Fick, 1993).

Trait Activation Theory

Trait Activation Theory links situational cues to behavior through the lens of trait relevance (Tett & Burnett, 2003; Tett et al., 2021), and posits that behavior is most likely to reflect latent personality traits when trait-relevant cues are present in the situation. In structured assessment contexts, individuals must interpret situational information and align their behavior accordingly, which has been shown to influence performance outcomes (Kleinmann et al., 2011). This process reflects the idea that traits are not expressed automatically but require the right kind of situation to become behaviorally visible.

For the present study, it is important to distinguish between basic objective situational information (*cues* in the sense of Rauthmann et al., 2015), and specific information that acts as a catalyst for inherent personality traits (Schäpers et al., 2022). Rooted in Trait Activation Theory (Tett et al., 2021), one driver of SJT response choice is the inclusion of situational cues that trigger behavior relevant to specific traits (Mussel et al., 2017). For these cues to effectively elicit diverse trait-related behavior in SJTs, situational strength should be moderate, ensuring that not all test-takers respond in the same way (Harris et al., 2016). Similarly, Marshall and Brown (2006) found that individuals high in a particular trait only require moderately strong situations to evoke a trait-relevant response. In contrast, strong situations, such as the aforementioned traffic light, provide little opportunity for variation in behavior, whereas moderate or weak situations, allow for greater variability in how individuals respond. Initial evidence suggests that this principle holds for SJTs, as

demonstrated by Schäpers et al. (2022) who showed that the inclusion of trait-relevant situational cues significantly influences trait-driven responses in these tests. In sum, Trait Activation Theory provides a bridge between objective situational cues and personality-driven behavior, particularly in contexts like SJTs where trait-relevant cues can be deliberately embedded to test this mechanism.

Simulation-Based Assessments in Personnel Selection

Theoretical insights into situation perception and construal are not only foundational in personality psychology but have also been increasingly applied to personnel selection, particularly in simulation-based assessment methods. Most prominent among these are assessment centers (ACs) and situational interviews, which are designed to simulate job-relevant situations (Lievens, 2017). These methods utilize situational cues to evoke behaviors and allow for the evaluation of job-relevant competencies. This evaluative process relies heavily on the test-taker's ability to correctly interpret situational cues, a skill referred to as the ability to identify criteria (ATIC, Kleinmann, 1993). Research has shown that this ability is predictive of success in both assessment centers and situational interviews, where the interpretation of situational demands is essential (Ingold et al., 2015). As SJTs also present scenarios that require situational reasoning, ATIC has often been considered relevant to their interpretation as well.

Situational Judgment Tests (SJTs) are considered low-fidelity simulations that provide short job-relevant scenarios followed by multiple behavioral response options (Motowidlo et al., 1990). Although intended to reflect workplace situations, research has questioned how situational these tests truly are in practice (Krumm et al., 2015b; Schäpers et al., 2020). SJTs may be understood as a combination of situation characteristics (that is, individual interpretation of objective circumstances), trait-relevant cues (that is, specific information driving the relevance of personality traits for behavioral choice) and a particular situational

strength (that is, the extent to which a situation can be understood and has to be interpreted). Thus, it is necessary to understand how individuals traverse the process from reading an SJT item to choosing a response option.

Ployhart (2006) introduced the predictor response process model, outlining four key phases involved in answering SJTs: comprehension, retrieval, judgment, and response selection. In the comprehension phase, individuals read, interpret, and make sense of the question or scenario presented in the SJT; that is, they process the relevant cue. The second phase, retrieval, involves accessing relevant knowledge and information stored in long-term memory about the given situation, that is, test-takers understand the situation's characteristics. In the third phase, judgment, test-takers use the retrieved information to form a decision or evaluation about the situation. Lastly, in the response selection phase, individuals review the available response options and select the one that aligns most closely with their judgment.

Similarly, the situated reasoning and judgement model by Grand (2020) proposes three steps of SJT responding: conditional reasoning (test-takers first read and interpret the situational scenario presented in the SJT, which involves understanding the context, demands, and key information provided in the item, or, again, understanding and interpreting the relevant cues), similarity judgement (test-takers assess the potential response options and draw on their past experiences to make judgments about which option best aligns with their understanding of the situation and its possible outcomes, that is, they give learned context to the objective information and evaluate the resulting situation characteristic), and preference accumulation (test-takers select the option they believe is the most appropriate based on their judgment formed in the previous phase, which draws from the understood characteristics of the situation and the individual's experiences).

In the tripartite model of SJT responding proposed by Martin-Raugh and Kell (2021), again three core cognitive processes comprise the SJT response. First, in the situation perception phase, test-takers interpret the situational information presented in the SJT item to grasp the context of the scenario, that is, interpret the objective information which results in the psychologically meaningful situation characteristic. Next, during goal formulation, individuals establish a desired outcome or objective by assessing the situational information, characteristics, and demands and aligning these with their personal values or expectations. During this, the goal is affected by how the situation is assessed in terms of its characteristics (that is, is it my goal to flee a threatening situation, or to make friends in a social setting?). Finally, in the response evaluation phase, test-takers assess the effectiveness of each response option in achieving the goal they have previously formulated.

In all of these process models, a consistent theme emerges: test-takers process objective information by integrating their personal context, resulting in a subjective interpretation of the situation, or what can be described as a situation characteristic. This interpretation forms the basis for how they evaluate response options and ultimately make their choice in response option. Given this fundamental role of situation characteristics in SJT response choice, the next important question is: How has previous research explored and assessed the various determinants of these situations?

Cues, Construal, and Item Components in SJTs

Research on the perception of simulation-based personnel selection procedures has shown that situation determinants as outlined above also have influence on these very specific types of situations: Brown et al. (2016) established that cues, characteristics, and classes are also featured in SJTs: Cues are often included in the context of the situation description, where the people present, the important actions, and relevant objects are delineated. In order to pick and choose a behavioral response options, individuals have to rely

Chapter 4: SJTs and Their Situation Construal

on their interpretation, that is, the characteristics of the given cues (Brown et al., 2016). By doing this, they can specify the situation information at hand: for example, an SJT item with the cues “another person, i.e. my supervisor, is present” and the class “workplace situation” can elicit very different behaviors depending on whether the main characteristics of the situation are Sociality (e.g., the supervisor wants to chat over a cup of coffee) or Duty (e.g. the supervisor wants to assign overtime work). While some interpretation might be already set by the information provided within the SJT item, most of the interpretation is still, like in real-life situations, done by the individual who shapes their response choice according to these situation determinants (Brown et al., 2016; Serfass & Sherman, 2013). This finding underscores the notion that understanding the situational demands is essential for effective performance in high-strength situations. When test-takers can accurately interpret the situational information, they are better equipped to align their behavior with the expectations of the situation, thus enhancing their performance. Ingold et al.'s research also draws parallels between situational interviews and SJTs, as both involve assessing responses to hypothetical scenarios. In tying these together, Whetzel et al. (2020) note that SJTs, like situational interviews, rely on the degree to which situational information can be understood, emphasizing the importance of cue interpretation for performance.

Krumm et al. (2015) conducted three studies in which they removed situational descriptions from SJTs to examine how strongly SJT performance was driven by situational information. Their results showed that for 43%-71% of items, removing the situation description did not significantly affect performance, indicating that these SJTs may measure general domain knowledge rather than situational judgment. Similarly, Jackson et al. (2017) found that most of reliable variance in SJT scores is attributable to candidate main effects, suggesting that a general judgment factor or within-person traits drive performance rather than variance related to specific situational information. This finding implies that SJT

performance may not strongly reflect situational specificity or dimension-specific judgments. Instead, SJTs may assess broader, generalized competencies such as judgment, challenging the assumption that SJTs primarily measure situation-specific responses. Schäpers et al. (2020) focused on both objective and subjective situational factors, exploring how the objective situation described in the SJT interacts with test-takers' subjective perceptions. The presence or absence of situational descriptions influenced the construct saturation of SJT scores, indicating the degree to which personality and cognitive abilities explain score variance. This ties into research on situational strength, as the manipulation of situational information alters the strength of cues presented to participants. On the other hand, Rockstuhl et al. (2015) found that in open-response SJTs, test-takers' ability to judge the situation (that is, their perception of objective situational cues and resulting characteristics) was a stronger predictor of job performance than their judgment of behavioral responses. This emphasizes the importance of how individuals construe situational elements over how they might choose to respond to them, reinforcing the traditional idea that accurate situation perception is key for SJT performance. While these studies differ in format and findings, they collectively suggest that the situational content of SJTs may not always reside where it is traditionally assumed to be, that is, in the item stem.

Kaminski et al. (2019) focused on the role of response options in SJTs, proposing that they serve as significant sources of information, thereby potentially reducing the need for detailed situational descriptions, which aligns with the findings by Krumm et al. (2015). Kaminski et al. (2019) highlight that response options carry inherent social desirability, which influences how test-takers make choices. This was tested across various SJT formats, showing that social desirability consistently correlates with response choices, suggesting that the desirability of response options is a key factor in performance. Freudenstein et al. (2020) build on the research on SJT response options and explored how test-takers perceive and

respond to situational information in SJTs using the DIAMONDS model (Rauthmann et al., 2014). Their research examined the role of situation construal, i.e. how individuals psychologically interpret the situations presented in SJT items, and found that it significantly contributes to SJT performance, whether or not explicit situation descriptions are provided: This suggests that in some SJTs, situational information is found primarily in the response options, and that test-takers draw their information from there instead of purely from the situation descriptions. The authors then extended this research by investigating the role of situational strength in SJT performance (Freudenstein, Schäpers, et al., 2023). While they found that situational strength did not moderate the relationship between personality traits and SJT performance, stronger situations (those with clearer cues and expectations) were associated with higher SJT scores. This indicates that the strength of the situation plays a role in shaping test-takers' performance in SJTs. Taken together, this raises the possibility that situational construal may be driven more by response options than by item stems, or at least that the two components contribute individually.

Mixed Evidence About Situation Construal in SJTs

As detailed above, some research suggests, situational determinants like DIAMONDS, situational strength, and social desirability perceptions all contribute to how test-takers process the situational information within SJTs, thereby shaping their response choices (Freudenstein, Schäpers, et al., 2023; Freudenstein et al., 2020; Kaminski et al., 2019). However, the research field is characterized by mixed evidence obtained from different studies with varying design. Hence, the next logical step is to explore how different conceptualizations of situation construal (such as DIAMONDS, situational strength, and social desirability) individually and collectively predict performance in SJT items, and which part of the SJT item (that is, the situation description or the response options) situation construal is most relevant in. Thus, in the current research we aim to investigate whether the

inclusion of trait-relevant cues impacts the predictive power of these situational characteristics in the item parts. We thusly manipulated seven versions of SJT items in our study to test inclusion and exclusion of different item parts and presence of trait-activating cues, and propose the following research questions and hypotheses:

RQ1: To what extent will different conceptualizations of situation construal (DIAMONDS, situational strength perceptions, social desirability perceptions) predict SJT item performance in full SJT items?

As delineated above, trait-relevant cues are generally thought to activate the expression of personality traits in behavior (Tett & Guterman, 2000; Tett et al., 2021). The presence of those cues should therefor increase the relevance of situation construal dimensions, as research by Brown et al. (2016) suggests that construal becomes more psychologically meaningful when trait-relevant information is contained in a situation. Incorporating the findings of Schäpers et al. (2022), we assume that the DIAMONDS, i.e. the operationalization of the psychological meaning of situations, should thus better predict SJT performance in conditions where trait-relevant cues point to trait-relevant behavior, and hypothesize:

H1: DIAMONDS will predict SJT item performance for all deconstructed SJT item versions that contain a trait-relevant cue better than for the respective item versions that do not contain a trait-related cue.

RQ2: To what extent will situation construal situational (in form of situational strength perceptions and social desirability perceptions) predict performance across different SJT item versions?

Methods

Sample

Participants were recruited via online panel Prolific (prolific.co) and compensated with 5€ for completing the questionnaire. As the questionnaire was in German, we applied screenings as per the panel rules and only contacted participants who were at least 18 years old and fluent in German. After excluding 45 individuals who asked for their data not to be used and 52 individuals who failed attention checks (two instructed response items, see Meade & Craig, 2012), our final sample included $N = 2.341$ participants ($\text{mean}_{\text{Age}} = 32.16$, $\text{SD}_{\text{Age}} = 11.77$, 43% female). Participants were gainfully employed (45.8% full time), with the majority working in service (26.23%) or education (23.58%) jobs. Participants were randomly assigned to one of seven groups, with group sample sizes of $N_1 = 324$, $N_2 = 303$, $N_3 = 340$, $N_4 = 308$, $N_5 = 306$, $N_6 = 318$, and $N_7 = 351$.

Study Design and Materials

Scales

To strike a balance between generalizability and participant burden, we administered a total of six SJT items stemming from two different SJTs, the Personal Initiative SJT (Bledow & Frese, 2009) and the Team Role Test (Mumford et al., 2008). Sample items from both SJTs are presented in Appendix C. The six items used in this study were adapted, translated, and tested by Schäpers et al. (2022) to contain a trait-relevant cue.

For each SJT item, we assessed test takers' situation construal. To this end, we administered an adapted version of the Situational Strength at Work Scale (Meyer et al., 2014) including the clarity, consistency, and constraints dimensions (three items each), as well as the Ultra Brief Measure for the Situational Eight DIAMONDS (one item per factor, Rauthmann & Sherman, 2015), and a social desirability one-item measure ("How relevant is

it for you to choose a socially desirable behavior in this situation?”). All items were presented in German using translation-backtranslation.

Conditions and Experimental Design

Participants were randomly assigned to one of seven different groups in which SJT items were presented as follows: (1) Full SJT item (including response options) with trait-relevant cue and context information in the item stem; (2) SJT item (including response options) with trait-relevant cue, but no context information in the item stem; (3) SJT item (including response options) with context information, but no trait-relevant cue in the item stem; (4) no SJT item stem (neither trait-related cues nor context information), only response options; (5) identical to (1), but without response options; (6) identical to (2), but without response options; (7) identical to (3), but without response options.

In each of the 7 conditions, we assessed test takers' situation construal per each SJT item. In versions 1 to 4, participants first responded to the SJT item before providing their situation construal. In versions 5 to 7, they first read the item text and provided their situation construal before seeing the response options and answering the item. This approach allowed us to assess how participants interpreted the SJT items' situation construal without being influenced by the response options. Additionally, two Subject Matter Experts (SMEs) assessed situation construal across all item variations as a comparison to participants' construal (first and last authors). Interrater reliability between SMEs was excellent, $ICC(A,2) = .96$, 95% CI [.95, .976], $F(754, 755) = 2.14$, $p < .001$. Accordingly, we averaged their ratings to form a single SME score for subsequent comparisons with the test-takers.

SJT Item Scoring

We applied distance scoring for SJT item evaluation, with each response option rated on a 7-point rating scale (ranging from 1 = *do not agree* to 7 = *fully agree*). If a response was deemed correct or indicative of a high trait standing, seven points were subtracted, meaning a

Chapter 4: SJTs and Their Situation Construal

rating of 7 was scored as 0 (no distance from the correct response), while a rating of 1 was scored as -6 (maximum distance). For incorrect or low-trait responses, one point was deducted. A rating of 1 was scored as 0, while a rating of 7 was scored as +6 (maximum distance). Neutral responses (neither correct nor incorrect) were assigned a midpoint deduction of four points. Final scores ranged from 0 (best) to 6 (worst), reflecting the distance from the ideal response (for a more detailed description of distance scoring, see Reznik et al., 2023).

Analytic Strategy

For situational strength (operationalized as clarity, consistency, and constraints) across the six SJT items, measurement invariance across the seven groups were conducted. For all items, we found metric invariance, but not scalar invariance. Thus, we refrain from mean difference analyses. For transparency, we provide mean difference comparisons for the DIAMONDS, but we refrain from interpreting as with one item per factor, measurement invariance testing cannot be conducted. Additionally, we compared participants' construal ratings to those of two SMEs to evaluate whether participants were able to meaningfully engage with the situational dimensions. For each item x construal combination, we computed mean values across participants and compared them to the SMEs' mean ratings. To quantify agreement, we calculated Pearson correlations across items, mean absolute errors (MAE) and root mean square errors (RMSE), and an intraclass correlation coefficient (ICC; two-way random effects, absolute agreement, average measures).

To evaluate the relevance of our aforementioned situation construal measures in predicting performance in the full item, we performed regression analyses on the item level. Only relevant predictors were then added into mixed-effects regression analyses across groups. Here, we deviate slightly from the preregistration

(https://osf.io/qf6j2/?view_only=11ae0f147f554756987b08524f38ba20), in that we initially

intended to include random intercepts (for participants in groups) and random slopes (for situation construal measures), but due to models failing to converge decided on reporting stable models with only random intercepts and fixed slopes when necessary. To test whether the influence of situation construal on item performance differed between groups containing versus not containing trait-relevant cues, we performed interaction regression analyses (see results section).

Results

Comparability of Situation Determinants Across Groups

For situational strength (operationalized as clarity, consistency, and constraints) across the six items, measurement invariance tests between the seven groups were conducted. For all items, we found at least metric invariance, but not scalar invariance (see Appendix A). Thus, we refrain from mean difference analyses.

Comparison with SMEs

Test-takers' mean construal ratings showed substantial convergence with SMEs. Across all items, the average deviation was $MAE = 1.37$ and $RMSE = 1.69$ scale points. The correlation between test-taker means and SME means was $r = .59$, indicating a moderately strong similarity in rank-order patterns. An $ICC(A,3) = .88$, 95% CI [.87, .90], $F(754, 976) = 8.80$, $p < .001$, further confirmed a high degree of agreement between SMEs and participants. Group-specific analyses yielded similar results, with Pearson correlations ranging between $r = .51$ and $.69$ and MAEs between 1.13 and 1.54. These findings suggest that participants were able to meaningfully construe the situational dimensions in ways largely consistent with expert judgments.

Research Question 1:

To examine the extent to which the different conceptualizations of situation construal predict SJT performance in full items, multiple regression analyses were conducted. Given that SJT scores represent deviation from the ideal response, negative regression coefficients indicate that a situation construal was associated with better performance in the SJT item (less deviation), whereas positive coefficients indicate that a situation construal was associated with worse performance in the SJT item (greater deviation from the ideal response).

The explanatory power of the models varied across items, with the highest proportion of variance explained in PI4 (8%), followed by PI1 (3%) and TRT7 (4%), all reaching statistical significance (for detailed results, see Table 1). In contrast, the model showed little to no predictive value for TRT8, with an adjusted R^2 close to zero. Only some DIAMONDS significantly predicted SJT item performance at all, and their effects differed substantially across items. Deception was, for example, associated with worse performance in TRT1, whereas Positivity and Sociality were linked to better performance in TRT7 and PI1, respectively.

Situational strength perceptions were also linked to SJT item performance only in certain cases. Clarity was associated with better performance in PI5, while Consistency was related to worse performance in PI4 and PI5. Additionally, Constraints were linked to worse performance in PI1. Social Desirability did not emerge as a significant predictor at all. Additional, not preregistered models focusing on DIAMONDS and Situational Strength separately showed differing predictive value. The DIAMONDS-only model performed comparably to the full model in some cases (PI1, PI4), whereas the Situational Strength-only model explained a larger proportion of variance in PI4 and PI5.

Hypothesis 1:

Interaction analyses examined whether DIAMONDS traits differentially predicted SJT performance depending on the presence of a trait-relevant cue. The results (for detailed results, see Table 2) varied across items, with significant interactions emerging for TRT1, TRT7, and PI1, while no meaningful effects were observed for TRT8, PI4, or PI5. For TRT1, Deception exhibited differential effects across groups, with a weaker association in Group 4 (vs. 1) and a stronger association in Group 7 (vs. 6). In TRT7, Positivity was more strongly associated with performance in Group 1 than in Group 4. Similarly, in PI1, Sociality had a greater impact on performance in Group 1 than in Group 4. No significant interaction effects were found for TRT8, PI4, or PI5, indicating that the presence of a trait-relevant cue did not moderate the relationship between DIAMONDS traits and SJT performance.

Chapter 4: SJTs and Their Situation Construal

Table 1

Regression Analyses Predicting Item Performance in the Full Item

Effect	Item TRT1		Item TRT7		Item TRT8		Item PI1		Item PI4		Item PI5	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	1.332	0.286	1.887	0.303	1.539	0.289	2.379	0.388	2.079	0.367	2.444	0.305
Duty (D)	0.020	0.034	-0.019	0.021	0.004	0.020	0.009	0.042	0.030	0.032	-0.036	0.030
Intellect (I)	-0.017	0.019	-0.001	0.025	0.034	0.026	-0.021	0.033	-0.045	0.028	0.017	0.026
Adversity (A)	-0.021	0.031	0.006	0.019	-0.029	0.020	-0.023	0.041	-0.020	0.035	-0.016	0.054
Mating (M)	0.026	0.031	-0.012**	0.031	-0.002	0.029	0.018	0.054	0.055	0.053	-0.024	0.045
pOsitivity (O)	-0.016	0.025	-0.105	0.039	-0.022	0.027	-0.037	0.046	0.017	0.048	-0.019	0.031
Negativity (N)	-0.016	0.023	-0.006	0.029	-0.030	0.022	0.001	0.036	-0.024	0.032	-0.024	0.026
Deception (De)	0.077**	0.029	-0.030	0.022	0.038	0.026	0.032	0.042	-0.074	0.043	0.024	0.047
Sociality (S)	0.013	0.026	0.051	0.030	0.040	0.035	-0.128**	0.039	-0.048	0.036	0.023	0.035
Clarity	0.037	0.026	-0.009	0.030	-0.030	0.026	-0.048	0.038	-0.042	0.037	-0.109**	0.031
Consistency	-0.018	0.028	0.003	0.030	0.016	0.028	0.083	0.045	0.161***	0.039	0.064	0.036
Constraints	-0.020	0.023	0.043	0.025	0.005	0.023	0.064*	0.032	0.052	0.036	0.037	0.034
Social Desirability	-0.022	0.022	-0.032	0.025	-0.050	0.026	-0.008	0.035	-0.019	0.030	0.000	0.028
Adjusted R ²	.007, $F(12, 324) = 1.218$, $p = .269$.04, $F(12, 324) = 2.168$, $p = .013$		-.002, $F(12, 324) = 0.96$, $p = .491$.03, $F(12, 324) = 1.929$, $p = .03$.08, $F(12, 324) = 3.402$, $p = .0001$.01, $F(12, 324) = 1.391$, $p = .169$	

Note. $N = 324$, SE = standard error. *** $p < .001$. ** $p < .01$. * $p < .05$. Due to the distance scoring method, direction of effects is inverted.

Table 2

Interaction Effects for DIAMONDS between Item Variations Containing and Not Containing Trait-Relevant Cues

Item / Group Comparison	Effect	Estimate	SE	p
TRT1				
1 : 4	Deception x Group	−0.024*	.009	.013
3 : 2		.013	.033	.683
5 : 6		.034	.029	.254
7 : 6		.0827**	.03	.0062
TRT7				
1 : 4	pOsitivity x Group	.111**	.042	.008
3 : 2		−.02	.05	.742
5 : 6		.034	.049	.479
7 : 6		.039	.05	.431
TRT8				
no significant DIAMONDS				
PI1				
1 : 4	Sociality x Group	.121**	.042	.004
3 : 2		−.023	.046	.609
5 : 6		−.054	.042	.188
7 : 6		−.034	.038	.375
PI4				
no significant DIAMONDS				
PI5				
no significant DIAMONDS				

Chapter 4: SJTs and Their Situation Construal

Note. Reference category is always the group containing the trait-relevant cue, SE = standard error. *** $p < .001$. ** $p < .01$. * $p < .05$. Due to the distance scoring method, direction of effects is inverted.

Chapter 4: SJTs and Their Situation Construal

Table 3

Mixed-Effects Regression for Situational Strength and Social Desirability

Effect	Item TRT1		Item TRT7		Item TRT8		Item PI1		Item PI4		Item PI5	
	Estimate	95 % CI	Estimate	95 % CI	Estimate	95 % CI	Estimate	95 % CI	Estimate	95 % CI	Estimate	95 % CI
Intercept	1.51***	[1.37, 1.64]	2.13***	[1.99, 2.27]	1.69	[1.52, 1.86]	2.24***	[2.04, 2.44]	2.54***	[2.35, 2.72]	2.58***	[2.38, 2.79]
Clarity	0.02	[-0.00, 0.03]	-0.00	[-0.02, 0.02]	0.01	[-0.01, 0.03]	-0.04***	[-0.06, -0.02]	-0.06**	[-0.08, -0.03]	-0.04**	[-0.06, -0.02]
Consistency	-0.00	[-0.02, 0.01]	-0.04***	[-0.06, -0.01]	-0.00	[-0.03, 0.02]	0.05***	[0.02, 0.08]	0.07**	[0.04, 0.09]	0.04**	[0.01, 0.06]
Constraints	-0.02*	[-0.03, -0.00]	-0.02	[-0.03, 0.00]	-0.02	[-0.04, -0.00]	0.01	[-0.01, 0.03]	-0.06**	[-0.09, -0.04]	-0.01	[-0.04, 0.01]
Social Desirability	-0.02**	[-0.04, -0.01]	-0.01	[-0.03, 0.01]	0.00	[-0.02, 0.02]	-0.06***	[-0.08, -0.04]	-0.01	[-0.04, 0.01]	-0.02*	[-0.04, -0.01]
Random effects												
σ^2	0.27		0.39		0.41		0.57		0.66		0.52	
τ_{00}	0.00		0.00		0.02		0.03		0.01		0.04	
ICC	0.01		0.00		0.04		0.04		0.02		0.07	
Marginal R ² / Conditional R ²	0.006 / 0.018		0.008 / 0.010		0.003 / 0.039		0.019 / 0.060		0.029 / 0.045		0.010 / 0.079	

Note. $N = 2341$, $N_{Groups} = 7$, CI = confidence interval, ICC = intraclass correlation coefficient. *** $p < .001$. ** $p < .01$. * $p < .05$. Due to the distance scoring method, direction of effects is inverted.

Research Question 2:

To evaluate the predictive value of situational strength and social desirability perceptions on SJT performance, we estimated mixed-effects models for each item. Due to convergence issues and indications of overparameterization (i.e., singular fit warnings), we followed recommendations by Bates et al. (2015) and Barr et al. (2013) and specified random intercept-only models, omitting random slopes to ensure model stability given the limited number of groups (for simplified model results, see Table 3, for the random-intercepts-random-slopes models, see Appendix B). Across items, clarity and consistency emerged as the most consistent predictors of performance. Specifically, greater perceived clarity was associated with better performance in three items, while higher consistency predicted better performance in three others. In contrast, both constraints and social desirability showed weaker and more sporadic relationships with performance. Random intercept variance was minimal across all models, and ICCs were low (ranging from .00 to .07), indicating that most of the variance occurred at the individual level. Likewise, marginal R^2 values were small ($\leq .03$), suggesting that situational strength and social desirability perceptions accounted for a limited portion of variance in SJT performance.

Discussion

The present study investigated the extent to which situation construal (operationalized via DIAMONDS dimensions, situational strength, and social desirability) predicted performance on different manipulations of SJT items. Results indicated that situation construal variables were predictive of performance in some cases. However, these effects were inconsistent across items. The proportion of variance explained by the models was minimal (Marginal $R^2 = .003-.027$), reflecting limited overall predictive utility. Item-specific effects were observed for select DIAMONDS dimensions, notably Deception, Sociality, and Positivity. Similarly, dimensions of situational strength, particularly Clarity and Consistency, emerged as significant predictors for a subset of items. Conversely, social desirability was

largely unrelated to performance across items. Interaction effects with trait-relevant cues were limited, as they occurred in only three of the six SJT items (TRT1, TRT7, and PI1), and also showed different directions, thus only partially supporting our hypothesis.

Theoretical Implications

As previously discussed, a long-standing question in the literature concerns the extent to which SJTs are situational in nature (Krumm et al., 2015; Schäpers et al., 2020). In the present study, although situation construal demonstrated some predictive relevance for SJT performance, situational effects were weak and inconsistent. We initially assumed a link between test-takers' situation construal and their response choice, i.e. their SJT performance. We did not find this link within our data, which challenges the current SJT response models as well as the situation construal model in itself. However, it is important to distinguish between two separate points: While in the present study participants' situation construal was not consistently associated with their item performance, this does not necessarily imply that test-takers did not experience the situations as psychologically real or meaningful. They may still have subjectively engaged with the situation, but this experience appeared to have limited predictive power for their response. This raises the possibility that the link between situational perception and behavioral response could be weaker than previously assumed, or that it may depend on other factors not examined in the present study (such as the salience of cues or the level of ambiguity).

The limited variance explained by situation construal variables raises questions about situation fidelity of SJTs (Motowidlo et al., 1990) and suggests that, for many test-takers, SJT items may not elicit the experience of "being in" a situation. SJTs have traditionally been considered as low-fidelity simulations (Motowidlo et al., 1990), but from our findings one could interpret this may be too low a threshold for reliably triggering meaningful construal-related responses. Indeed, higher-fidelity formats, such as multimedia (Rockstuhl et al., 2015) or immersive virtual environments, may be better suited to elicit proper situational immersion.

Chapter 4: SJTs and Their Situation Construal

Recent proposals for standardized state assessments (Freudenstein, Schulze, et al., 2023) also support the idea that more structured and clear representations may enhance the understanding of ‘being in’ a situation. Thus, while we cannot rule out that participants engaged with the situational information on a subjective level, our results suggest that this engagement may not systematically inform their behavioral choices. Future research should aim to disentangle situational experience from its behavioral correlates, for example by combining construal assessments with immersion ratings or real-time processing measures.

Our results also challenge the general applicability of existing SJT response process models, such as the predictor response process model (Ployhart, 2006), the situated reasoning and judgment model (Grand, 2020), and the tripartite model (Martin-Raugh & Kell, 2021), all of which assume that test-takers engage in substantial situation construal as a core step in the response process. While these models describe situation interpretation as a fundamental component of SJT responding, the present findings suggest that situation construal influences performance only in a limited and somewhat unsystematic manner. This discrepancy suggests that existing models may overstate the relevance and extent of situational reasoning processes within SJTs, particularly when situational information (Saucier et al., 2007) are weak, ambiguous, or not salient enough to prompt meaningful interpretation. Instead, test-takers may engage in situation construal selectively, depending on the interpretability of the information provided. This notion is in line with recent discussions on the importance of standardized state assessments that offer clearer, more structurally defined situational information (Freudenstein, Schulze, et al., 2023)

Importantly, the presence of trait-relevant cues did not substantially moderate the relationship between situation construal and performance. This finding stands in contrast to expectations based on Trait Activation Theory (Tett et al., 2021), which posits that trait-relevant situational cues should elicit trait-expressive behavior and, by extension, stronger and more differentiated performance. While it remains plausible that trait-relevant cues triggered

Chapter 4: SJTs and Their Situation Construal

some level of trait activation, the absence of a robust and systematic interaction effect in SJTs suggests that such activation was insufficient to truly alter the way test-takers construed the situation. This resonates with the distinction of Rauthmann et al. (2014) between cue recognition and the experience of a situation as psychologically meaningful. Based on our results, simply embedding a trait-relevant cue does not appear sufficient to make an SJT item psychologically "situational" in a way that fundamentally guides judgment and response processes. Thus, the assumption that trait-relevant cues inherently strengthen situation-trait linkages in SJTs may need reconsideration. While our findings suggest that the mere presence of trait-relevant cues does not consistently enhance the predictive power of situation construal, this stands in contrast to some previous studies that did find systematic effects. For example, Schäpers et al. (2022) demonstrated that trait-relevant cues can influence trait-driven responses under certain conditions. Taken together, these findings suggest that the role of trait cues may be more context-dependent than previously assumed, and that it is possibly additionally influenced by cue salience, item structure, or test-taker strategies.

Finally, prior research has consistently identified social desirability as a key influence on SJT performance (Brown & Martin-Raugh, 2024; Kaminski et al., 2019; Leeds, 2018). However, in the present study, social desirability was largely unrelated to SJT item performance. This finding is noteworthy because, unlike many previous studies, we examined the relationship between social desirability perception and performance at the item level, allowing for a more fine-grained analysis of how perceived social appropriateness influences behavior. One possible interpretation is that the testing context in our study (that is, low-stakes, anonymous, and online) did not activate participants' motivation to respond in socially desirable ways. Another possibility is that our item content lacked sufficient ambiguity or moral tension to make social desirability salient, as has been observed in prior research.

Practical Implications

The present findings offer several implications for the design and development of SJTs. First, the inconsistent predictive power of situation construal variables highlights the need for structure in item design. Up until now, focus in SJT development has been on including as much trait information as possible to counteract the hot mess that is SJT construct validity (McDaniel et al., 2016), thereby potentially losing sight of the situational aspects. Test developers might aim to construct items that present clear and interpretable situational contexts, thereby reducing ambiguity and minimizing cognitive burden. Rather than assuming that test-takers will naturally construe complex scenarios as psychologically meaningful, developers might explicitly embed situational information in ways that facilitate comprehension. Careful attention might also be given to where and how situational information are presented within the item structure. Whether situational information is embedded in the item stem or the response options may influence how test-takers process and interpret the situation (Freudenstein et al., 2020). Future SJT development should systematically evaluate how different item components affect situation perception and response behavior.

Moreover, the present results suggest that SJTs may function more as frame-of-reference tests than as genuine simulations of real-world situations (Lievens & Sackett, 2017). That is, they may offer only contextual framing without fully immersing test-takers in a simulated experience. This distinction has important design implications: more deliberate selection of response options, potentially through single-response SJTs, may be necessary to enhance interpretability. Event-reconstruction approaches, as proposed by Grube et al. (2008), could support the development of more situationally rich items. Similarly, adopting more standardized and structured item formats, as advocated by Freudenstein, Schulze, et al. (2023), may enhance situational clarity and improve the coherence of SJTs.

Future research

The present findings highlight several new possibilities for future research. One particularly relevant avenue concerns the assessment of mean-level differences in situational perception across different item versions. Although the present study could not examine such differences due to a lack of scalar measurement invariance, the violation of invariance itself might be theoretically meaningful (Protzko, 2025): it suggests that participants may have construed situations differently depending on the presence or absence of cues and context information. In future studies, focus should be put on using within-subject designs to circumvent between-group comparability issues (Greenwald, 1976).

Also, there is a need to broaden the focus of SJT research from trait saturation to situational saturation. Historically, the field has disproportionately emphasized trait activation and individual differences, which might have been at the expense of a deeper understanding of the situational features that shape test-taker behavior. Given the limited and inconsistent effects of trait-relevant cues observed in the present study and also the still lacking construct validity despite of almost a decade of focused research since identifying it as a hot mess (McDaniel et al., 2016), future research should more directly examine the characteristics of situations themselves, including how they are perceived, construed, and operationalized within SJTs.

In pursuit of this aim, the use of automated item generation and artificial intelligence presents a promising avenue. Previously, researchers were successful in using large language models for creating psychometrically sound SJTs (Krumm et al., 2024; Schäpers et al., 2024). These technologies could enable easy systematic manipulation of situational features across large item pools, allowing researchers to isolate and examine the specific situational determinants that influence SJT performance.

Additionally, future studies should explore the concept of situation "classes" as proposed by Rauthmann et al. (2015). Rather than focusing solely on discrete situational

Chapter 4: SJTs and Their Situation Construal

features, examining higher-order categories of situations such as workplace conflicts, collaborative tasks, or ethical dilemmas may yield insights into functional equivalencies across different contexts. Investigating how individuals perceive and respond to these broader situation classes could enhance theoretical models of situation perception and inform the development of more ecologically valid and generalizable SJT frameworks.

Limitations

Several limitations of the present study should be noted. First, although the experimental multigroup design allowed for manipulation of item components, the item pool was limited to six items drawn from two instruments, which was necessary to reduce participant burden but arguably constrains the generalizability of our findings. The item-specific effects observed in our analyses may not extend to other domains, constructs, or response formats. Future research using broader and more diverse item sets is needed to establish whether the patterns reported here are robust across a wider range of SJT items.

Second, although we aimed to comprehensively assess situation construal, our measurement approach relied on ultra-brief instruments (once again in order to limit participant burden). The DIAMONDS were each represented by a single item, as was social desirability perception, which might make these measures susceptible to measurement error and limited construct representation (Fisher et al., 2016). This limitation is particularly relevant for social desirability, where we observed no predictive relationship with SJT performance despite prior research identifying it as an important factor (Brown & Martin-Raugh, 2024; Kaminski et al., 2019; Leeds, 2018). Future research should consider using more extensive social desirability measures and also experimentally manipulate impression management demands to clarify under which conditions social desirability exerts an influence. Similarly, multi-item assessments of DIAMONDS dimensions could help better capture the richness of situational construal, if participant burden is acceptable.

Chapter 4: SJTs and Their Situation Construal

Finally, although the study design allowed for detailed experimental control, findings are based on a single dataset using a custom item set. As such, conclusions about situation construal processes in SJTs should be interpreted in light of this scope. Replication across independent samples, item types, and applied testing contexts, as well as across full SJTs is essential to determine the extent to which these findings generalize across assessment settings.

References

- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *Preprint arXiv:1506.04967*. <https://doi.org/10.48550/arXiv.1506.04967>
- Bledow, R., & Frese, M. (2009). A 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, M., & Martin-Raugh, M. (2024). Exploring the role of social desirability in situational judgment tests. *PsyArXiv Preprints*. <https://doi.org/10.31234/osf.io/wxqt3>
- 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>
- Fischer, D. G., & Fick, C. (1993). Measuring social desirability: Short forms of the Marlowe-Crowne social desirability scale. *Educational and Psychological Measurement*, 53(2), 417–424. <https://doi.org/10.1177/0013164493053002011>
- Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology*, 21(1), 3–23. <https://doi.org/10.1037/a0039139>
- 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>
- Freudenstein, J.-P., Schäpers, P., Reznik, N., Stolte, T., & Krumm, S. (2023). The influence of situational strength on the relation of personality and situational judgment test

performance. *International Journal of Selection and Assessment*.

<https://doi.org/10.1111/ijsa.12444>

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*, 73(4), 669–700.

<https://doi.org/10.1111/peps.12385>

Freudenstein, J.-P., Schulze, J., Schäpers, P., Mussel, P., & Krumm, S. (2023). Standardized state assessment: A methodological framework to assess person-situation processes in hypothetical situations. *European Journal of Psychological Assessment*.

<https://doi.org/10.1027/1015-5759/a000794>

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>

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology*, 60(5), 773–794. <https://doi.org/10.1037//0022-3514.60.5.773>

Grand, J. A. (2020). A general response process theory for situational judgment tests. *Journal of Applied Psychology*, 105(8), 819–862. <https://doi.org/10.1037/apl0000468>

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83(2), 314–320. <https://doi.org/10.1037/0033-2909.83.2.314>

Grube, A., Schroer, J., Hentzschel, C., & Hertel, G. (2008). The event reconstruction method: An efficient measure of experience-based job satisfaction. *Journal of Occupational*

and Organizational Psychology, 81(4), 669–689.

<https://doi.org/10.1348/096317907X251578>

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>

Hough, L. M., & Oswald, F. L. (2008). Personality Testing and Industrial-Organizational Psychology: Reflections, Progress, and Prospects. *Industrial and Organizational Psychology-Perspectives on Science and Practice*, 1(3), 272–290. <https://doi.org/10.1111/j.1754-9434.2008.00048.x>

Ingold, P. V., Kleinmann, M., König, C. J., Melchers, K. G., & Van Iddekinge, C. H. (2015). Why do situational interviews predict job performance? The role of interviewees' ability to identify criteria. *Journal of Business and Psychology*, 30, 387–398. <https://doi.org/10.1007/s10869-014-9368-3>

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), 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>

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology*, 78(6), 988–993. <https://doi.org/10.1037/0021-9010.78.6.988>

Chapter 4: SJTs and Their Situation Construal

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011).

A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review*, *1*(2), 128–146.

<https://doi.org/10.1177/2041386610387000>

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), 399–416. <https://doi.org/10.1037/a0037674>

Krumm, S., Thiel, A. M., Reznik, N., Freudenstein, J.-P., Schäpers, P., & Mussel, P. (2024).

Creating a psychological test in a few seconds: Can ChatGPT develop a psychometrically sound situational judgment test? *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000878>

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>

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., & Sackett, P. R. (2017). The effects of predictor method factors on selection

outcomes: A Modular Approach to Personnel Selection Procedures. *Journal of Applied Psychology*, *102*(1), 43–66. <https://doi.org/10.1037/ap10000160>

Marshall, M. A., & Brown, J. D. (2006). Trait aggressiveness and situational provocation: A

test of the traits as situational sensitivities (TASS) model. *Personality and Social Psychology Bulletin*, *32*(8), 1100–1113. <https://doi.org/10.1177/0146167206288488>

Martin-Raugh, M. P., & Kell, H. J. (2021). A process model of situational judgment test responding. *Human Resource Management Review*, *31*(2), 100731.

<https://doi.org/10.1016/j.hrmr.2019.100731>

Chapter 4: SJTs and Their Situation Construal

- 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., 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>
- 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>
- Meyer, R. D., & Dalal, R. S. (2009). Situational strength as a means of conceptualizing context. *Industrial and Organizational Psychology*, 2(1), 99–102. <https://doi.org/10.1111/j.1754-9434.2008.01114.x>
- 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. <https://doi.org/https://doi.org/10.4324/9780203763643>
- 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/10.1037/0033-295X.102.2.246>

Chapter 4: SJTs and Their Situation Construal

- 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>
- 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., Schäpers, P., Freudenstein, 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>
- 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.
- Protzko, J. (2025). Invariance: What does measurement invariance allow us to claim? *Educational and Psychological Measurement, 85*(3), 458–482. <https://doi.org/10.1177/00131644241282982>
- 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. (2015). Measuring the situational eight DIAMONDS characteristics of situations: An optimization of the RSQ-8 to the S8*. *European Journal of Psychological Assessment, 31*(1), 1–10. <https://doi.org/10.1027/1015-5759/a000246>

Chapter 4: SJTs and Their Situation Construal

- 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>
- 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>
- Reznik, N., Krumm, S., Freudenstein, J. P., Heimann, A. L., Ingold, P., Schäpers, P., & Kleinmann, M. (2023). Does understanding what a test measures make a difference? On the relevance of the ability to identify criteria for situational judgment test performance. *International Journal of Selection and Assessment, 32*(2), 210–224.
<https://doi.org/https://doi.org/10.1111/ijsa.12458>
- 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), 479–504. <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, 87*, 103963.
<https://doi.org/10.1016/j.jrp.2020.103963>
- Schäpers, P., Krumm, S., Lievens, F., Freudenstein, J.-P., Schulze, J., & König, C. J. (2022). Situation descriptions in situational judgment tests: A matter of trait activation? *European Association of Work and Organizational Psychology*, Glasgow, Scotland.

Chapter 4: SJTs and Their Situation Construal

- Schäpers, P., Reznik, N., Rapp, J., Schöne, T., & Heinemann, H. (2024). *Eine KI-unterstützte SJT-Entwicklung zur situativen Messung von Social Entrepreneurship*. 53rd Congress of the German Psychological Society (DGPs), Vienna, Austria.
- Serfass, D. G., & Sherman, R. A. (2013). Personality and perceptions of situations from the Thematic Apperception Test. *Journal of Research in Personality, 47*(6), 708–718. <https://doi.org/10.1016/j.jrp.2013.06.007>
- Tett, R. P., & 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>
- Tett, R. P., Toich, M. J., & Ozkum, S. B. (2021). Trait activation theory: A review of the literature and applications to five lines of personality dynamics research. *Annual Review of Organizational Psychology and Organizational Behavior, 8*, 199–233. <https://doi.org/10.1146/annurev-orgpsych-012420-062228>
- Whetzel, D. L., Sullivan, T. S., & McCloy, R. A. (2020). Situational judgment tests: An overview of development practices and psychometric characteristics. *Personnel Assessment and Decisions, 6*(1), 1. <https://doi.org/10.25035/pad.2020.01.001>

Appendix A

Measurement Invariance

Table A1

Results for Measurement Invariance Testing of Situational Strength Across Groups

Item	Invariance Level	χ^2 (<i>df</i>)	CFI	RMSEA	Δ CFI	Δ RMSEA	Decision
TRT1	Configural	302.90 (168)	0.972	0.049			Acceptable
	Metric	349.47 (204)	0.970	0.046	-.002	-.003	Acceptable
	Scalar	476.00 (240)	0.951	0.054	-.019	-.008	Invariance not supported
TRT7	Configural	436.75 (168)	0.944	0.069			Acceptable
	Metric	482.72 (204)	0.942	0.064	-.002	-.005	Acceptable
	Scalar	664.46 (240)	0.911	0.073	-.031	-.009	Invariance not supported
TRT8	Configural	382.63(168)	0.960	0.062			Acceptable
	Metric	426.19 (204)	0.959	0.057	-.001	-.005	Acceptable
	Scalar	582.40 (240)	0.937	0.065	-.022	-.008	Invariance not supported
PI1	Configural	301.60 (168)	0.979	0.049			Acceptable
	Metric	346.43 (204)	0.978	0.046	-.001	-.003	Acceptable
	Scalar	532.42 (240)	0.954	0.060	-.023	-.015	Invariance not supported
PI4	Configural	419.58 (168)	0.950	0.067			Acceptable
	Metric	475.76 (204)	0.946	0.063	-.004	-.004	Acceptable
	Scalar	639.90 (240)	0.921	0.071	-.025	-.007	Invariance not supported
PI5	Configural	416.54 (168)	0.957	0.067			Acceptable
	Metric	465.32 (204)	0.955	0.062	-.002	-.005	Acceptable
	Scalar	583.54 (240)	0.941	0.065	-.014	-.004	Invariance not supported

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; Δ CFI and Δ RMSEA represent changes relative to the previous, less constrained model. Chi-square values are based on the maximum likelihood estimator. Scalar invariance is not supported where Δ CFI exceeds .010.

Appendix B

Regressions

Table B1

Mixed-Effects Regression for Situational Strength and Social Desirability

Effect	Item TRT1		Item TRT7		Item TRT8		Item PI1		Item PI4		Item PI5	
	Estimate	95 % CI	Estimate	95 % CI	Estimate	95 % CI	Estimate	95 % CI	Estimate	95 % CI	Estimate	95 % CI
Intercept	1.50***	[1.37, 1.63]	2.11***	[1.91, 2.32]	1.68***	[1.46, 1.90]	2.26***	[1.99, 2.53]	2.47***	[2.15, 2.79]	2.57***	[2.35, 2.79]
Clarity	0.02	[-0.00, 0.04]	-0.00	[-0.02, 0.02]	0.01	[-0.02, 0.04]	-0.04**	[-0.07, -0.01]	-0.06***	[-0.09, -0.03]	-0.04*	[-0.07, -0.01]
Consistency	-0.00	[-0.04, 0.03]	-0.04**	[-0.06, -0.01]	-0.01	[-0.04, 0.03]	0.05*	[0.01, 0.08]	0.07*	[0.02, 0.12]	0.04	[-0.00, 0.07]
Constraints	-0.02	[-0.03, 0.00]	-0.01	[-0.04, 0.01]	-0.02	[-0.05, 0.00]	0.01	[-0.03, 0.05]	-0.05*	[-0.09, -0.02]	-0.01	[-0.05, 0.02]
Social Desirability	-0.02**	[-0.04, -0.01]	-0.01	[-0.03, 0.01]	0.00	[-0.02, 0.03]	-0.05***	[-0.08, -0.03]	-0.01	[-0.04, 0.02]	-0.02	[-0.04, 0.00]
Random effects												
σ^2	0.27		0.39		0.40		0.57		0.65		0.51	
τ_{00}	0.00		0.04		0.05		0.09		0.13		0.05	
τ_{11} Group x Clarity	0.00		0.00		0.00		0.00		0.00		0.00	
τ_{11} Group x Consistency	0.00		0.00		0.00		0.00		0.00		0.00	
τ_{11} Group x Constraints	0.00		0.00		0.00		0.00		0.00		0.00	
τ_{11} Group x Social Desirability	0.00		0.00		0.00		0.00		0.00		0.00	
ρ_{01} Clarity	0.76		-1.00		-0.19		-0.35		-1.00		-0.18	
ρ_{01} Consistency	-0.48		0.08		-0.49		-0.30		-0.51		0.87	
ρ_{01} Constraints	-0.22		-0.99		-0.25		-0.43		-0.35		-0.97	
ρ_{01} Social Desirability	0.03		-0.37		-0.04		-0.31		-0.43		-0.68	
Marginal R^2	.006		.008		.003		.019		.027		.010	

Note. $N = 2341$. CI = confidence interval. *** $p < .001$. ** $p < .01$. * $p < .05$. Due to the distance scoring method, direction of effects is inverted.

Appendix C

SJT Items

Sample Items:

TRT8:

Sie sind Mitglied eines sechsköpfigen Produktionsteams. Ihre Arbeitsbelastung ist angestiegen und Sie werden daher ein zusätzliches Teammitglied einstellen. Ihr Team ist für die Herstellung von Waren im Textilbereich verantwortlich. Im Allgemeinen ist das Unternehmen gut aufgestellt und verfügt über verschiedene Produktionsstandorte sowie ein weit gefächertes Sortiment. Sie haben fünf Bewerber interviewt und sind gegenwärtig in einem Meeting mit Ihrem Team, in dem diskutiert wird, welcher Bewerber ein Einstellungsangebot bekommen soll. Zwei Teammitglieder diskutieren kontrovers, welcher Bewerber am geeignetsten für die Stelle erscheint und können sich nicht auf eine Lösung einigen. Es liegt eine große Anspannung in der Luft. Was würden Sie tun?

- A. Auch wenn die beiden Kollegen sich nicht auf eine Lösung einigen, beteiligen Sie sich nicht an der Diskussion, weil klar ist, dass diese in einem Streit ausartet und Sie den Streit nicht eskalieren lassen wollen.
- B. Sie lassen die Teammitglieder die Situation diskutieren, da Sie nicht in langandauernde Meinungsverschiedenheiten hineingezogen werden wollen.
- C. Sie bleiben still, bis ein Teammitglied seine Ansichten darlegt, denn so gewinnen Sie einen besseren Eindruck davon, wie das Team über das Thema denkt.
- D. Da Sie merken, dass sich die beiden Teammitglieder nicht auf eine Lösung einigen können, intervenieren Sie und schlagen vor, dass das Team zunächst die Kriterien für die Auswahl eines Bewerbers aufschreiben sollte, bevor es beginnt, über irgendeinen bestimmten Bewerber zu diskutieren.

P11:

Sie arbeiten nun bereits seit einiger Zeit für ein großes Familienunternehmen in Frankfurt am Main. In der Abteilung, in der Sie arbeiten, wird ein neues Computerprogramm installiert. Sie fühlen sich unsicher im Umgang mit dem Programm. Sie und einige Ihrer Kollegen machen immer wieder Fehler, die sie alle viel Zeit kosten. Aus Zeit- und Kostengründen wurde bei der Einführung des Programms keine ausführliche Schulung organisiert. Was würden Sie tun?

- A. Ich organisiere eine interne Schulung, in der erfahrene Kollegen Ihr Wissen über das Programm weitergeben. So können die zeitraubenden Fehler reduziert werden, die mir und anderen unerfahrenen Kollegen immer wieder passieren.
- B. Ich mache Überstunden und bleibe wenn nötig noch länger im Büro. So können die zeitraubenden Fehler reduziert werden, die mir und anderen unerfahrenen Kollegen immer wieder passieren.
- C. Ich erarbeite mir selbst das fehlende Wissen aus Büchern nach Dienstschluss. So können die zeitraubenden Fehler reduziert werden, die mir und anderen unerfahrenen Kollegen immer wieder passieren.
- D. Ich rege mich nicht weiter darüber auf, bleibe gelassen und fokussiere mich weiterhin auf meine Arbeit. So können die zeitraubenden Fehler reduziert werden, die mir und anderen unerfahrenen Kollegen immer wieder passieren.

Chapter 5

Discussion

General Discussion

This final chapter provides an integrative discussion of the findings presented in the preceding chapters, situating them within the current theoretical landscape of SJT research. The overarching aim of this dissertation was to examine the question of how situational SJTs are, after all, and specifically how test-takers construe SJT items as psychologically meaningful situations. Drawing on person-situation-interaction theory (Funder, 2006, 2016; Mischel, 1977) and current SJT process models (Grand, 2020; Martin-Raugh & Kell, 2021; Ployhart, 2006), I introduced a working model of SJT responding that conceptualizes SJT performance as the product of interactions between person properties (traits such as trait social desirability), item properties (e.g., situation descriptions, response options, presence of trait-relevant cues), and the psychological interpretation of the situation (situation construal via situational strength, Situational Eight DIAMONDS, the Ability To Identify Criteria (ATIC), social desirability and effectiveness perceptions).

Across three empirical papers comprising four studies, different facets of this model were examined. Chapter 2 investigated the role of ATIC as a person-side indicator of situation construal. Chapter 3 focused on which components of an SJT item (namely, item stems or response options) drive construct and criterion validity. Chapter 4 provided a direct test of the role of situation construal in SJT responding by operationalizing construal via DIAMONDS, situational strength, and social desirability perception ratings across various item versions.

In the present chapter, the findings of these studies are integrated and critiqued with regard to the working model. Key implications for SJT theory, development, and validation are derived regarding the nature of SJT situationality, the relevance (or irrelevance) of situation construal for test performance, and the shifting conceptualization of SJTs from behavioral

Chapter 5: Discussion

simulations toward generalized, structurally contextualized assessments. This is followed by a discussion of methodological and conceptual limitations, and implications for future research.

Chapter 2: Situational Judgment Tests and the Ability to Identify Criteria (ATIC)

ATIC refers to individuals' capacity to understand what is being measured in a given assessment situation (Kleinmann et al., 2011), and was proposed as an explanatory construct for individual differences in performance on context-rich tasks, initially for assessment center exercises and situational interviews. Chapter 2 extended ATIC research to SJTs, as prior research on the subject has been inconsistent across SJT presentation forms. SJTs are likewise framed as contextualized assessments and theoretically assumed to require perception and understanding of situational demands (Motowidlo et al., 1990; Wolcott et al., 2021). The question at hand was whether ATIC, as a proxy for a specific form of situation construal, predicts SJT performance in a meaningful and generalizable way.

Two studies addressed this question. In Study 1, ATIC was measured using open-ended questions following 55 SJT items that had been drawn from multiple pre-existing SJTs covering a wide range of constructs (e.g., integrity, empathy, teamwork). The link between ATIC and SJT performance was modeled on the item level. In Study 2, a whole SJT (Freudenstein, Remmert, et al., 2020) was used, and participants were randomly assigned to a high- or low-incentive condition to test the influence of motivational processes on ATIC's predictive utility. Across both studies, ATIC was assessed via Subject Matter Expert (SME) ratings of participants' ability to correctly infer what each item was evaluated.

Overall, the results were not encouraging for the predictive power of ATIC in SJTs. In Study 1, ATIC predicted SJT performance only weakly. Additionally, most of the variance in ATIC scores was attributed to the constructs being measured by the item, rather than to stable individual differences: some constructs such as honesty-humility were just easier to identify than others (such as personal initiative). This raises concerns about whether ATIC, in the context

of SJTs, truly reflects an individual ability rather than an item feature, which would situate it on the side of the situation rather than the person in the working model. Study 2 yielded a similarly unexpected pattern. ATIC predicted SJT scores more strongly under low-incentive than under high-incentive conditions, which is surprising considering the fact that prior ATIC research found substantial relationships specifically in high-stakes selection processes (Ingold et al., 2015). Our findings suggest that, if ATIC is at all predictive for SJT performance, higher motivation might crowd out its utility: it is possible that in high-stakes SJT contexts knowledge and performance motivation take over in driving SJT performance, while in low-stakes SJT contexts the typical motivation to perform falls away, leaving “only” ATIC as a performance driver.

These results raise theoretical doubts about the role of ATIC in explaining SJT performance. Prior research had suggested that ATIC was a key individual driver of performance in simulation-based formats (Ingold et al., 2015; Kleinmann et al., 2011), but such findings may not generalize to the low-fidelity, text-based SJTs used in this study. In contrast to structured interviews and assessment centers, the SJTs employed here may not have required the same level of situation construal. Instead, participants may rely on surface-level knowledge or response heuristics that reduce the diagnostic role of situation construal (Freudenstein, Schäpers, et al., 2020; Kaminski et al., 2019; Krumm et al., 2015). In this light, ATIC (as operationalized in chapter 2) may not reflect a stable individual difference variable, but rather a context-dependent aspect of item difficulty. That is, ATIC might be better conceptualized not as a person-side variable (as assumed in the working model), but as an item-side feature indicating how easily a given item’s evaluative construct can be inferred. This interpretation is further supported by the observation that item-level ATIC scores did not consistently align with performance variance, which calls into question whether this type of construal is even activated in this format. Another important implication of this study is that the presumed simulation logic

Chapter 5: Discussion

of SJTs may be overstated. If ATIC plays a limited or item-specific role, it suggests that test-takers do not meaningfully construe SJT items as psychologically rich situations. Instead, they may respond based on pre-existing social knowledge, perceived item difficulty, or the relative plausibility and desirability of response options. This weakens the theoretical assumption embedded in trait activation theory (Tett & Burnett, 2003), namely, that trait expression depends on recognizing trait-relevant situational cues. While this idea may hold in richer assessment formats, it was not supported here. More broadly, these results add to the debate on SJT processing, supporting the view that SJTs operate less as simulations and more as contextualized knowledge tasks.

Chapter 3: Situational Judgment Tests and Their Components

In chapter 3, to exhaustively test the different parts of SJT items, I employed a within-subjects longitudinal design in which participants completed four versions of eight SJT items: response options, response options in a randomized list (with instructions varying between behavioral, effectiveness, and social desirability), open-ended responses to the item stem, and the full, unmodified item. Items were drawn from two SJTs targeting conscientiousness and personal initiative. The study assessed how well each item version predicted the full item, corresponding self-report measures, and self- and supervisor-rated job performance. Across analyses, response options were consistently the strongest predictor of full item responses. In most items, adding situation stems or open responses contributed little incremental variance. Moreover, randomized response options showed some predictive utility for self- and supervisor-rated job performance, especially under behavioral and knowledge instructions, but again did not outperform the response options together. Regarding construct validity, randomized response options with behavioral or effectiveness instructions showed the strongest, albeit inconsistent, correlations with self-report measures, while other versions demonstrated limited

Chapter 5: Discussion

construct alignment. Open-ended responses were the weakest predictor of full item responses and only sporadically predicted external criteria.

The findings of this study carry significant theoretical implications. Most critically, they again challenge the assumption that SJTs function as simulations requiring situation construal. Response options rather than contextual information as provided in the item stem were the strongest predictor of full item performance, calling into question both simulation conceptualization (Brown et al., 2016; Motowidlo et al., 1990; Motowidlo et al., 1997) and key tenets of SJT process models (Grand, 2020; Martin-Raugh & Kell, 2021; Ployhart, 2006). Within the working model of SJT responding outlined in chapter 1, this suggests a shift in focus: from situation construal as a mechanism mostly on the side of the person to properties of the item itself, especially the clarity and plausibility of response options. Rather than engaging in the psychological interpretation of item information, participants may rely on stable heuristics (Krumm et al., 2015; Naemi et al., 2016) or trait-aligned response tendencies (Kaminski et al., 2019) when choosing response options.

Trait social desirability (TSD), originally tested in the present study as a moderator between item versions and their corresponding self-report questionnaires, can instead be understood as an underlying influence in SJTs across item versions. TSD was found to act as a default response strategy (Kaminski et al., 2019) aligning with recent research that treats it as a structural factor in SJT responding (Brown & Martin-Raugh, 2024), and was highly predictive of corresponding constructs in the present study, while failing to consistently emerge as a significant moderator. These findings suggest that TSD may operate as a structural component of SJT responding rather than merely influencing it.

All of this ties into the findings in chapter 2, where ATIC was assumed to be a person-side situation construal component, but turned out to be mostly influenced by item variance. These findings support a reconceptualization of SJTs as structurally contextualized

Chapter 5: Discussion

measurement tools in which situationality emerges from the information within item components (particularly response options) rather than from elaborated construal processes based on the item stem-

Chapter 4: Situational Judgment Test Components and Their Individual Situation Construal Properties

Chapter 2 showed how ATIC (conceptualized as a person-side variable reflecting situation construal) had limited predictive utility for SJT performance, suggesting that construal processes may play a different role in SJTs as presumed. Chapter 3 further extended this logic by showing that response options demonstrated the most consistent construct validity, highlighting their relevance in SJT solving processes. Against this backdrop, chapter 4 systematically investigated the role of situation construal as defined in chapter 1 via the SCM (Funder, 2016), by examining the relationship between construal and item performance across different SJT item components. For this, I tested how various operationalizations of situation construal (DIAMONDS, situational strength, and social desirability perception) predict SJT performance in full and deconstructed items, and whether this depends on the presence of trait-relevant cues. The study employed an experimental between-subjects design in which participants completed one of seven versions of six SJT items, systematically varying item stem versions (including contextual information, trait-relevant cues, or both) and the presence of response options. Participants also rated the items on DIAMONDS, situational strength, and social desirability. Across regression and mixed-effects models, situation construal variables demonstrated minimal and inconsistent predictive value for item performance. While specific construal dimensions (e.g., Sociality, Deception, Consistency) were weakly linked to item performance in some isolated cases, no generalizable pattern emerged. Interaction effects with trait-relevant cues were similarly rare and item-specific, with no consistent amplification of

Chapter 5: Discussion

construal-performance links. Social desirability perceptions, hypothesized to influence item processing, were largely unrelated to performance.

The implications of these results are, again, somewhat extensive. Most importantly, they call into question the assumption that SJTs responding is dependent on situation construal. Within the framework of the working model, situation construal was originally conceptualized as a person-driven mechanism (Funder, 2016) that informed and influenced SJT response choice. However, the findings did not support this at all, and instead directly challenge the response process models that underlie the working model (Grand, 2020; Martin-Raugh & Kell, 2021; Ployhart, 2006).

Our findings suggest two possible things: either situation construal does simply not inform or influence SJT responding, or SJT items are not understood, that is, construed, as situational. The first interpretation follows the logic initially proposed by Krumm et al. (2015), in that the situation (that is in the case of chapter 4, the item stem presenting with trait-relevant cues and/or contextual information) is not that important for SJT solving, and that alternative strategies such as relying on general domain knowledge (Fan et al., 2016; Lievens & Motowidlo, 2016), effectiveness cue-based reasoning (Leeds, 2018), or socially desirable responding (Kaminski et al., 2019) are more likely to be in effect. Although this seems counterintuitive at first when social desirability perception was not related to item performance in chapter 4, it ties these results in well with the findings of chapter 3, where trait social desirability was found to attenuate construct validity in some SJT items. This suggests that socially desirable responding may operate less through explicit construal of “socially desirable content” and more as a default heuristic or dispositional strategy that bypasses explicit construal.

The second interpretation that SJT items are not construed as situations gains support from the observed failure of trait-relevant cues to consistently affect the predictive validity of

Chapter 5: Discussion

situation construal variables, which contradicts prior research (Schäpers et al., 2022). Although trait-relevant cues were systematically embedded in the item stems to facilitate trait activation, they did not consistently interact with construal operationalizations in predicting item performance. This goes against a central premise of Trait Activation Theory (Tett & Burnett, 2003), which posits that trait expression in behavior is dependent on the perception of trait-relevant situational cues. The findings therefore either suggest a failure of cue detection altogether, or a failure of cue detection to elicit a psychologically meaningful situation construal. The distinction between cue detection and construal has been suggested in recent situation perception research (Kuper et al., 2024), which posits that situational features must not only be noticed but also cognitively integrated and imbued with relevance in order to influence behavior. In the present study, the weak and inconsistent associations between construal variables and SJT performance across item versions indicate that such cognitive integration found in real-life situations may not reliably occur in our text-based, low-fidelity SJTs. Rather than functioning as simulations that engage the same situation construal processes as the “real deal” (that is, real-life situations), these items may fall below the cognitive threshold necessary to activate them.

These conclusions resonate with results from chapters 2 and 3. Where chapter 2 found that ATIC was largely driven by item characteristics, and chapter 3 showed response options to be the primary driver of SJT performance, chapter 4 now demonstrates that direct construal assessments are not robust predictors of item performance.

Implications for SJT Theory: From Simulation to Structurally Contextualized Assessments

The working model proposed in chapter 1 was developed to synthesize prevailing theoretical assumptions from person-situation-interaction research (Funder, 2016), Trait

Chapter 5: Discussion

Activation Theory (Tett & Burnett, 2003) and current SJT response process models (Grand, 2020; Martin-Raugh & Kell, 2021; Ployhart, 2006), and it conceptualized SJT performance as the result of dynamic interactions between person variables, situation construal, and item properties. The results of the present dissertation, however, largely failed to support this architecture. Instead, they suggest a structurally and not dynamically contextualized idea of SJT responding in which situation construal is less central and performance is attributed more directly to properties of the items themselves.

Across chapters 2-4, mostly item properties accounted for performance and validity, whereas construal measures showed weak, inconsistent, or null predictive value. This effectively reverses the assumed flow in the working model. Instead of a process in which test-takers read a scenario, construe its meaning, and then select a response, results suggest that, as argued by Melchers and Kleinmann (2016), the situational meaning is already encoded in the response content itself. Freudenstein, Schäpers, et al. (2020) supported this view by showing that SJT response options can contain more DIAMONDS-based situational information than their corresponding stems, and that this embedded information can sustain the construal-performance link even without the stem. From this perspective, SJTs appear “contextualized” not because test-takers actively simulate behavior in situations, but because contextual cues are structurally embedded within the item, and especially within the response options. This finding directly challenges process models that suggest consecutive stages (e.g., understanding -> interpreting -> response choice). Instead of interpretive processing, the present results suggest that test-takers primarily rely on recognizing and applying cues embedded in the item. This also contradicts the pathway assumed in Trait Activation Theory (Tett & Burnett, 2003; Tett & Guterman, 2000), which holds that trait expression requires both the presence and recognition of trait-relevant cues. Even when such cues were explicitly embedded in stems (chapter 4), they did not interact with construal measures to influence responding. This suggests that cue salience

Chapter 5: Discussion

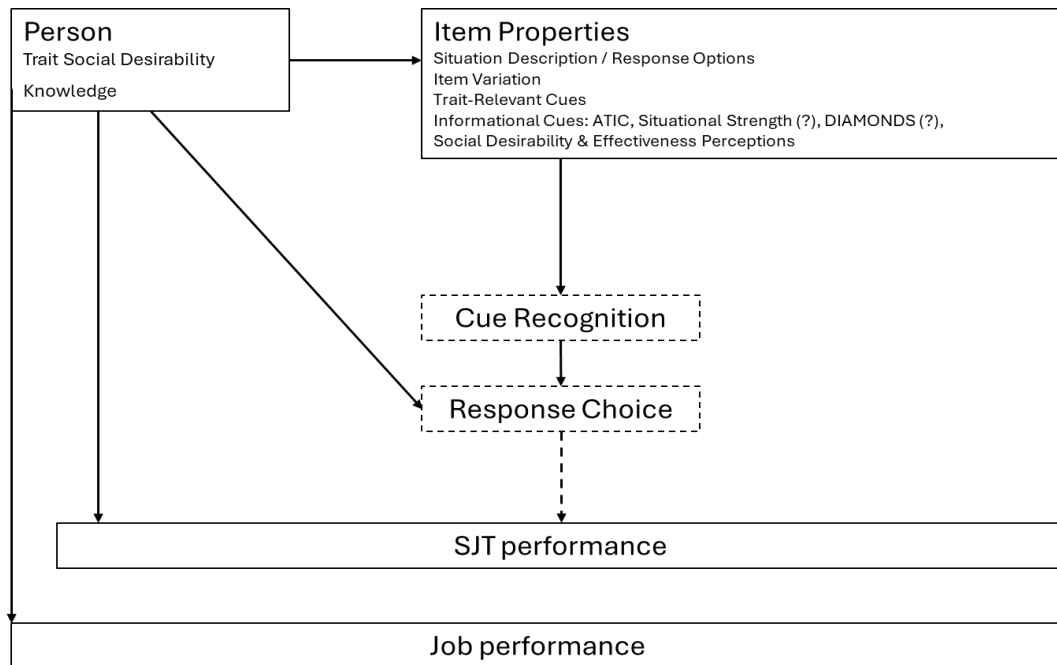
alone is insufficient unless it triggers a meaningful psychological representation, which was not observed in the present data.

Viewed from this perspective, SJTs should more accurately be described as structurally contextualized assessments: instruments in which situationality arises from embedded informational cues and linguistic framing within the item architecture. These cues are structural affordances and not emergent from dynamic person-situation interaction. As such, SJT performance may reflect the ability to recognize and apply knowledge to these embedded cues rather than the ability to simulate behavior in psychologically meaningful situations. This cue-based view aligns with empirical demonstrations that removing scenario descriptions does not impair validity (Krumm et al., 2015; Schäpers et al., 2020), that trait social desirability can act as primary driver of responding (Kaminski et al., 2019), and that effectiveness judgments can be made from response options alone without context comprehension (Leeds, 2018).

In this revised conceptualization, SJTs are structurally, not dynamically contextualized. Distinguishing cue recognition from situation construal extends prior research and suggests that, in typical text-based SJTs, situation construal may not be psychologically relevant at all. This reframing moves SJTs even further away from the simulation paradigm and toward a cue- and knowledge-based model, in which performance reflects cue recognition and knowledge application rather than dynamic trait activation through simulated scenarios (see Future

Figure 2

Updated Working Model of SJT Responding



Note. Visual representation of model integration across person and item properties.

Research. Regarding the aforementioned working model, an updated version (see Figure 2) would still include the operationalizations used in chapters 2-4, but they would be understood primarily as item-embedded properties processed by the test-taker, rather than as elements of a dynamic interaction.

Implications for SJT Design and Validation

The present findings raise substantive questions about the construct validity of SJTs. If psychological situation construal (as per the present operationalization) is not meaningfully linked to trait expression or item performance, then a core assumption of SJT design is undermined. Specifically: what do SJTs measure, and how should they be interpreted?

Chapter 5: Discussion

Traditionally, construct validity arguments have leaned on the simulation logic: if SJTs present realistic work scenarios and these scenarios activate trait-relevant behavior, then item responses should reflect the target construct (Motowidlo et al., 1990; Weekley & Ployhart, 2006). The present results challenge this logic in typical low-fidelity (that is, text-based) SJT formats. Instead, they indicate that performance is driven primarily by the structural features of the response options, not by active psychological engagement with situational content. These implications are most directly applicable to low-fidelity formats and may not fully generalize to higher-fidelity, multimedia, or interactive SJTs, where situational richness could possibly increase psychological meaningfulness (Cucina et al., 2015; Golubovich et al., 2017; Rockstuhl et al., 2015). However, evidence from situational interview and assessment center research shows that accurately recognizing situational demands, be it measured directly or indirectly, remains an important predictor of job performance (Ingold et al., 2015; Jansen et al., 2010; Speer et al., 2014). This suggests that the diminished role of construal observed here may be partly a function of format and design choices in the SJTs studied, rather than a universal feature of all SJTs.

Hands-on, these insights imply several design priorities. First, structural functioning and response option quality should be prioritized over narrative elaboration. Design efforts should focus on ensuring that the cues embedded in item components (wherever they are included, but specifically within the response options) are construct-relevant, diagnostically clear, and discriminable. This includes optimizing option plausibility and embedded trait-relevant information, as well as aligning both stems and options along a construct continuum where applicable (Thiel et al., 2024). Validation should demonstrate that these embedded cues reliably elicit trait-relevant judgments across populations, as diagnostic function is more important than perceived realism. Construct-driven SJT development can help achieve this by explicitly

Chapter 5: Discussion

targeting a single trait or compound trait, increasing internal consistency and reducing construct contamination (Holtrop & Oostrom, 2025).

At the same time, realism and comprehensibility need to be balanced: While the findings of prior studies and, to an extent, of this dissertation, suggest that cue comprehensibility may outweigh narrative realism for validity, realism (that is, fully phrased whole items) can still influence applicant burden and perceived fairness in SJTs (Kepes et al., 2025). The optimal balance between realism and comprehensibility will depend on the test's intended purpose (e.g., high-stakes selection vs. developmental feedback).

Also, there are concerns implied regarding fairness, accessibility, and subgroup differences. A stronger focus on structural cues could reduce subgroup performance differences by minimizing language complexity and culturally specific references in SJT items (Prasad et al., 2016). However, differences in background knowledge may still influence cue interpretation. Validation should therefore include differential item functioning (DIF) analyses to assess fairness (Oliveri et al., 2016).

Finally, systematic scenario design remains crucial. The absence of a consistent construal-performance link suggests that many current SJT items may lack the internal coherence, comprehensibility, or ecological salience needed to trigger meaningful construal. Melchers and Kleinmann (2016) caution that when the requirement for situational judgment is too limited (for example, in overly transparent items), systematic variance relevant to job performance may be lost, thereby potentially reducing criterion-related validity. One approach to address this is the standardized state assessment (Freudenstein et al., 2023) which replaces rich but ambiguous scenarios with brief, clearly bounded descriptions (e.g., "Your new neighbor invites you to their birthday party. When you arrive, you realize that you don't know any other guests") that carry well-defined trait implications. This reduces interpretive variance and allows for a tighter link between item content and intended construct. Another approach is to

systematically evaluate items in advance with representative samples to ensure consistent construal and that the response options exhaustively cover behavioral possibilities. However, such overarching evaluation is resource-intensive and assumes broad interpretability across diverse populations.

In sum, SJT development should move from intuitive, language-based realism to systematic structural standardization. If situation construal cannot be relied upon as a consistent driver of performance, then the emphasis should shift to ensuring that embedded cues (wherever in the item) are consistently interpretable, construct-relevant, and diagnostic. Under this approach, the “situational” in “situational judgment test” becomes a property of the item’s structural design, rather than of the test-taker’s dynamic interpretation.

What’s Left To Do with SJTs? Limitations and Future Research

Limitations

In addition to the specific limitations noted within the three studies, several overarching limitations need to be addressed. All three empirical chapters relied exclusively on text-based, low-fidelity SJTs. While such SJTs are widely used in research and practice because of their ease of development and administration (Motowidlo et al., 1990), they differ from higher-fidelity formats (e.g., video-based or interactive SJTs). They lack the non-verbal, dynamic, and affective contents that may increase immersion (Lievens et al., 2005; Lievens & Sackett, 2006) and potentially strengthen situation construal processes. It therefore remains unclear whether the patterns observed here (the limited role of situation construal and the dominant role of response options) would generalize to richer formats in which situational information is conveyed through multiple channels. Notably, item fidelity or cue strength was not systematically manipulated across a continuum. While chapter 4 included trait-relevant cue

Chapter 5: Discussion

manipulations based on pretested items by Schäpers et al. (2022), the expected cue-construal-performance link could not be consistently found. This suggests that, although cue salience was established in earlier research, these cues may not have been sufficiently diagnostic within the present deconstructed item formats. As such, the absence of construal effects in chapter 4 may not reflect the irrelevance of trait-relevant cues per se, but rather a disconnect between embedded cues and meaningful interpretation in this specific context.

This concern also relates to construct validity and its assumed relationship to Trait Activation Theory (Tett & Burnett, 2003). This theory assumes that trait-relevant cues in SJT items activate latent traits, resulting in measurable trait expression. Within the present research, self-report trait measures and SJT responses did not consistently converge: the pattern of prediction across item versions and analyses was inconsistent and generally weak. This problem is likely due to the notoriously low construct validity of SJTs (McDaniel et al., 2016), and further compounded by the fact that, with the exception of the second study in chapter 2 (which used the full teamwork SJT), the research relied primarily on isolated or component-level item formats, not full SJTs. In chapters 3 and 4, item stems and response options were deliberately decomposed to isolate structural effects, which was an approach suitable for process examination but not necessarily for construct validity examination. This aligns with the concern raised by McDaniel et al. (2016), who argue that SJT items are typically heterogeneous at the item level, meaning that individual items often load on multiple constructs, which leads to difficulties in establishing factor structure or discriminant validity (see also Kasten & Freund, 2016; and MacKenzie et al., 2010). Thiel et al. (2024) found similar results: in an AI-supported reanalysis of the data collected in chapter 3, they showed that although the intended construct was usually strongly or even most strongly represented in the items, other constructs were also featured. This ambiguity poses a specific challenge with regards to Trait Activation Theory, as multiple traits could be activated by a single item, leading to construct con- and diffusion. Taken

together, these findings underscore the difficulty of interpreting SJT responses as unambiguous indicators of trait expression. SJT items often contain multi-construct content, even when designed with a primary target trait in mind. This within-item heterogeneity complicates the attribution of SJT performance to a specific construct and may explain the weak convergence with self-report traits observed in the present research. Without behavioral process data or stronger construct-level alignment across items, distinctions between the relevance of trait activation, general domain knowledge, or surface-level reasoning strategies remain theoretically and empirically unresolved.

Additionally, there are possible concerns about the specific measurement of situation construal. Situation construal, central to the working model presented in chapter 1, was assessed using several theory-based indicators: ATIC, DIAMONDS, situational strength, and social desirability perception. While these were aligned with theoretical consensus (Funder, 2006, 2016), they may have offered only partial coverage of the construct. First, both the DIAMONDS and situational strength scales were administered in an ultra-short version, or in abbreviated and adapted formats, which was done to minimize participant burden. These choices may have reduced measurement precision and may have contributed to attenuated effect sizes in construal-performance relationships. Second, measurement invariance issues further limit interpretability. In chapter 4, the situational strength scale was found to be non-invariant across experimental groups, indicating that participants may have interpreted or used the scale differently depending on item presentation. For the DIAMONDS dimensions, measurement invariance could not be tested at all, as each dimension was represented by only one item. As a result, it is unclear whether the observed weak and inconsistent associations reflect a true lack of relationship or are artifacts of lacking measurement properties. Finally, as data collection was done via online panels, the studies did not incorporate process-tracing methods (like eye-tracking, think-aloud protocols) that could have provided live indicators of situation

Chapter 5: Discussion

interpretation and decision-making. Without such methods, the role of construal remains inferred from self-report ratings, limiting conclusions about whether, when, and how test-takers engaged in processing. Together, these limitations allow the possibility that the weak relationships observed between construal and performance may reflect or be compounded by instrument-level constraints rather than a genuine irrelevance of construal during SJT responding.

Finally, there are some overarching issues with sampling and contextual constraints. All three studies were conducted using online, low-stakes samples. While these samples were relatively large, gender-balanced, and demographically somewhat diverse (meaning that they probably surpass the convenience samples often found in psychological research), they do not represent high-stakes applicant populations typically encountered in personnel selection. In chapter 2, the introduction of a high-incentive condition led to higher self-reported motivation and slightly higher SJT scores for some item versions. However, these incentives did not change the relationship between situation construal and performance. This suggests that while motivation can increase under higher stakes, it may not necessarily engage or disengage the construal processes assumed by the working model. Because the present data come from low-stakes contexts, the findings should be generalized to applied selection settings (where SJTs are prominently used, see Webster et al., 2020) with caution. In real-world assessments where outcomes carry significant consequences, test-taking behavior could differ in important ways, including greater depth of engagement and increased strategic responding (Anglim et al., 2018). Without data from such contexts, the role of construal and cue use under true high-stakes conditions remains an open question.

Future research

Chapter 5: Discussion

With the present dissertation, several questions remain unanswered, or even stem from the results and/or the previously discussed limitations themselves. To address these questions, I propose a three-study plan:

The first study should test whether the weak role of construal is specific to text-based formats or whether it generalizes to higher-fidelity SJTs. A between-subjects experiment could therefore directly compare text-based and video-based SJTs using parallel items, i.e. items with the same situational information. Construal would be measured via DIAMONDS (but using the full four items per facet-scale, Rauthmann et al., 2014). AI-assisted item generation (Krumm et al., 2024) could ensure functional equivalence across text and video versions, systematically matching linguistic and visual cue presentation, and validating interpretability in pretesting pipelines (Schwarz, 2025). The contribution of such a study would be to clarify whether the irrelevance of construal found here is an artifact of the specific low-fidelity text-based SJTs used in this dissertation, or a general property of SJTs across modalities. The design of this study would also allow for both panel admission and laboratory assessment, with the latter allowing for additional construal measurement such as think-aloud assessment.

A second study could disentangle the processes of cue detection, construal, and trait activation. Chapter 4 showed that embedding and omitting trait-relevant cues in item stems did not yield consistent cue-construal effects, raising the possibility that participants noticed cues but did not integrate them meaningfully. Future work could manipulate cue salience parametrically (e.g. from weak to moderate to strong) across stems and response options of an SJT, then measure cue detection directly through recognition tasks, construal through DIAMONDS, and trait expression through behavioral validation tasks such as gamified versions of the same SJT content (Ohlms et al., 2024). This would allow a more thorough test of Trait Activation Theory in SJTs. AI-supported item generation could efficiently produce parallel items with controlled cue variations. The working hypothesis is that weak cues will still

Chapter 5: Discussion

be detected, but only strong cues will translate into meaningful construal and observable trait-relevant expression. The contribution of such a study would be to provide direct empirical evidence on whether and how cues progress from detection to construal and finally to trait activation, addressing a central theoretical question raised by the present work.

The third study should examine whether SJT responding and construal is construct-specific. In the present studies, items targeted multiple constructs including big five personality constructs, teamwork, team roles, and personal initiative, but analyses did not systematically test across construal measures whether construal mattered more for some constructs than for others. Chapter 1 showed initial evidence that ATIC was a construal operationalization that was construct-dependent but did not specifically test this for full SJTs measuring different constructs. Prior SJT research indicates that SJTs targeting interpersonal constructs, such as teamwork or leadership, tend to show stronger criterion-related validity than those assessing more procedural or rule-based domains (Christian et al., 2010). It therefore seems reasonable to assume that interpersonal SJTs rely more heavily on dynamic situation construal, while procedural SJTs are solved primarily through cue recognition and the application of domain knowledge. A within-subjects design could test this by having participants complete multiple construct-driven SJTs. For each construct, construal, cue use, and performance could be measured and validated against a multi-trait multi-method matrix including self-reports, peer ratings, and supervisor assessments. Here again, AI-assisted item generation (Krumm et al., 2024; Schäpers et al., 2024), or, more economically, AI-assisted content analysis of preexisting items (Thiel et al., 2024), could help build construct-specific item pools that maximize correspondence to the intended trait while minimizing cross-construct contamination. The contribution of this work would be to determine whether the role of construal is uniform across psychological domains or whether it varies systematically by construct, thus refining the theoretical scope of Trait Activation Theory in SJT contexts.

Chapter 5: Discussion

Taken together, these studies would extend the present work by testing (a) whether construal effects differ in higher-fidelity formats, (b) whether cue detection, construal, and trait activation can be separated empirically, and (c) whether construal plays a differential role depending on the construct domain. By combining systematic manipulations of format, cue salience, and construct domain with the affordances of AI-supported item development, analysis, and pretesting, future research could provide a decisive answer to the problem raised by the present dissertation: if at all, under what conditions, and for which constructs, do SJTs function as situational measures rather than structurally contextualized knowledge tests?

Conclusion

This dissertation set out to examine how “situational” SJTs really are and whether and how performance depends on situation construal. Across four empirical studies in three papers, the findings were consistent in their inconsistency: construal measures showed weak and inconsistent predictive relationships, whereas structural item features (specifically the response options) proved more important. In the examined SJTs, situationality thus appears less a matter of dynamic interpretation and more a property embedded in item design. Theoretically, this reframes SJTs as structurally contextualized assessments rather than low-fidelity simulations. In this view, performance reflects recognition of embedded cues, plausibility judgments, or social desirability tendencies rather than the activation of traits through situation construal. This aligns with prior evidence that stems can be removed without impairing validity (Krumm et al., 2015; Schäpers et al., 2020) and with critiques highlighting item heterogeneity and construct diffusion (McDaniel et al., 2016; Thiel et al., 2024). For practice, these insights shift design priorities away from narrative realism and toward systematic structural standardization

Chapter 5: Discussion

(Freudenstein et al., 2023). Response option quality, cue understandability, and construct alignment are more critical than elaborate scenario descriptions.

With respect to the working model proposed in chapter 1, the present findings suggest that its emphasis on dynamic construal processes overestimates their role in SJT performance. While the integrated concepts of person, situation, and item remain useful, they appear to operate less as interacting processes than as properties embedded within the item design and recognized by test-takers. The updated model therefore places item features as the primary locus of situationality within SJTs.

In conclusion, SJTs remain valuable, usable predictors of work-relevant outcomes, but not for the reasons long assumed. Their validity rests not on dynamic situation construal but on the structural embedding of specific properties (like response options, trait-relevant cues, response framing) within items. The central answer to this dissertation's guiding question “how situational are SJTs?” is therefore: SJTs predict performance not because test-takers actively construe situations, but because situational meaning is already built into their structure, making them situational by design, but not by interpretation.

References

- Anglim, J., Bozic, S., Little, J., & Lievens, F. (2018). Response distortion on personality tests in applicants: comparing high-stakes to low-stakes medical settings. *Advances in Health Sciences Education, 23*(2), 311-321. <https://doi.org/10.1007/s10459-017-9796-8>
- Brown, M., & Martin-Raugh, M. (2024). Exploring the Role of Social Desirability in Situational Judgment Tests. <https://doi.org/https://doi.org/10.31234/osf.io/wxqt3>
- 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>
- Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational Judgement 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>
- Cucina, J. M., Su, C. W., Busciglio, H. H., Thomas, P. H., & Peyton, S. T. (2015). Video-based Testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment, 23*(3), 197-209. <https://doi.org/10.1111/ijsa.12108>
- Fan, J. Y., Stuhlman, M., Chen, L. J., & Weng, Q. X. (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>
- Freudenstein, J. P., Remmert, N., Reznik, N., & Krumm, S. (2020). *English translation of the Teamwork Situational Judgment Test (SJT-TW). Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. <https://doi.org/10.6102/zis285>

Chapter 5: Discussion

- 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/https://doi.org/10.1111/peps.12385>
- Freudenstein, J.-P., Schulze, J., Schäpers, P., Mussel, P., & Krumm, S. (2023). Standardized state assessment: A methodological framework to assess person-situation processes in hypothetical situations. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000794>
- 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/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/https://doi.org/10.1177/0963721416635552>
- Golubovich, J., Seybert, J., Martin-Raugh, M., Naemi, B., Vega, R. P., & Roberts, R. D. (2017). Assessing Perceptions of Interpersonal Behavior with a Video-Based Situational Judgment Test. *International Journal of Testing*, 17(3), 191-209. <https://doi.org/10.1080/15305058.2016.1194275>
- Grand, J. A. (2020). A general response process theory for situational judgment tests. *Journal of Applied Psychology*, 105(8), 819. <https://doi.org/https://doi.org/10.1037/apl0000468>
- Holtrop, D., & Oostrom, J. K. (2025). Consequences of Adding Context in Personality Assessment. *Current Opinion in Psychology*, 102092. <https://doi.org/https://doi.org/10.1016/j.copsyc.2025.102092>
- Ingold, P. V., Kleinmann, M., König, C. J., Melchers, K. G., & Van Iddekinge, C. H. (2015). Why do situational interviews predict job performance? The role of interviewees' ability to identify criteria. *Journal of Business and Psychology*, 30(2), 387-398. <https://doi.org/https://doi.org/10.1007/s10869-014-9368-3>

Chapter 5: Discussion

- Jansen, A., Melchers, K., König, C., Kleinmann, M., Brändli, M., Fraefel, L., & Lievens, F. (2010). Candidates who correctly identify situational demands show better performance. 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA,
- 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/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(3), 230-240. <https://doi.org/10.1027/1015-5759/a000250>
- Kepes, S., Keener, S. K., Lievens, F., & McDaniel, M. A. (2025). An integrative, systematic review of the situational judgment test literature. *Journal of management*, 51(6), 2278-2319. <https://doi.org/https://doi.org/10.1177/01492063241288545>
- Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review*, 1(2), 128-146. <https://doi.org/https://doi.org/10.1177/2041386610387000>
- 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), 399-416. <https://doi.org/10.1037/a0037674>
- Krumm, S., Thiel, A. M., Reznik, N., Freudenstein, J.-P., Schäpers, P., & Mussel, P. (2024). Creating a psychological test in a few seconds: Can ChatGPT develop a psychometrically sound situational judgment test? *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000878>

Chapter 5: Discussion

- Kuper, N., Ernesti, L., Modersitzki, N., Phan, L. V., & Rauthmann, J. F. (2024). Distinctions without differences? Effects of instruction sets for situation characteristic ratings. *European Journal of Psychological Assessment*.
<https://doi.org/https://doi.org/10.1027/1015-5759/a000867>
- Leeds, J. P. (2018). Applying cognitive acuity theory to the development and scoring of situational judgment tests. *Behavior Research Methods*, 50(6), 2215-2225.
<https://doi.org/10.3758/s13428-017-0988-1>
- Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, 90(3), 442-452. <https://doi.org/10.1037/0021-9010.90.3.442>
- 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., & 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>
- MacKenzie, W. I., Ployhart, R. E., Weekley, J. A., & Ehlers, C. (2010). Contextual Effects on SJT Responses: An Examination of Construct Validity and Mean Differences Across Applicant and Incumbent Contexts. *Human Performance*, 23(1), 1-21,
<https://doi.org/10.1080/08959280903400143>
- Martin-Raugh, M. P., & Kell, H. J. (2021). A process model of situational judgment test responding. *Human Resource Management Review*, 31(2), 100731.
<https://doi.org/https://doi.org/10.1016/j.hrmr.2019.100731>

Chapter 5: Discussion

- 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>
- Mischel, W. (1977). On the future of personality measurement. *American Psychologist*, 32(4), 246. <https://doi.org/10.1037/0003-066X.32.4.246>
- 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., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology*. (pp. 241-260). Davies-Black Publishing.
- Naemi, B., Martin-Raugh, M., & Kell, H. (2016). SJTs as Measures of General Domain Knowledge for Multimedia Formats: Do Actions Speak Louder Than Words? *Industrial and Organizational Psychology-Perspectives on Science and Practice*, 9(1), 77-83. <https://doi.org/10.1017/iop.2015.121>
- Ohlms, M. L., Melchers, K. G., & Kanning, U. P. (2024). Playful personnel selection: The use of traditional versus game-related personnel selection methods and their perception from the recruiters' and applicants' perspectives. *International Journal of Selection and Assessment*, 32(3), 381-398. <https://doi.org/10.1111/ijsa.12466>
- Oliveri, M. E., Ercikan, K., Lyons-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied Measurement in Education*, 29(1), 17-29. <https://doi.org/10.1080/08957347.2015.1102913>

Chapter 5: Discussion

- 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., & Weekley, J. A. (2006). 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. 345-350). Lawrence Erlbaum Associates Publishers.
- Prasad, J. J., Showler, M. B., Schmitt, N., Ryan, A. M., & Nye, C. D. (2016). Using Biodata and Situational Judgment Inventories across Cultural Groups. *International Journal of Testing*, 1-24. <https://doi.org/10.1080/15305058.2016.1218338>
- 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). <https://doi.org/10.1037/a0037250>
- 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>
- 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*, 87, 103963. <https://doi.org/10.1016/j.jrp.2020.103963>
- Schäpers, P., Krumm, S., Lievens, F., Freudenstein, J.-P., Schulze, J., & König, C. J. (2022). *Situation Descriptions in Situational Judgment Tests: A Matter of Trait Activation?* European Association of Work and Organizational Psychology, Glasgow, Scotland.

Chapter 5: Discussion

- Schäpers, P., Reznik, N., Rapp, J., Schöne, T., & Heinemann, H. (2024). *Eine KI-unterstützte SJT-Entwicklung zur situativen Messung von Social Entrepreneurship*. 53rd Deutsche Gesellschaft für Psychologie (DGPs) Congress, Vienna, Austria.
- Schwarz, A. (2025). *Automatic Item Generation for Non Cognitive Constructs* [Report in preparation]. Hogrefe Publishing Group
- Speer, A. B., Christiansen, N. D., Melchers, K. G., König, C. J., & Kleinmann, M. (2014). Establishing the Cross-Situational Convergence of the Ability to Identify Criteria: Consistency and Prediction Across Similar and Dissimilar Assessment Center Exercises. *Human Performance*, 27(1), 44-60.
<https://doi.org/10.1080/08959285.2013.854364>
- Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, 88(3), 500-517.
<https://doi.org/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/https://doi.org/10.1006/jrpe.2000.2292>
- Thiel, A. M., Reznik, N., & Krumm, S. (2024, September 15-21). *But Wait, There is More! Über die in SJTs angesprochenen Konstrukte* 53rd Deutsche Gesellschaft für Psychologie (DGPs) Congress, Vienna, Austria.
- Webster, E. S., Paton, L. W., Crampton, P. E., & Tiffin, P. A. (2020). Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education*, 54(10), 888-902. <https://doi.org/https://doi.org/10.1111/medu.14201>
- 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. 1-10). Lawrence Erlbaum Associates Publishers.

Chapter 5: Discussion

Wolcott, M. D., Lobczowski, N. G., Zeeman, J. M., & McLaughlin, J. E. (2021). Does the ability to identify the construct on an empathy situational judgment test relate to performance? Exploring a new concept in assessment. *Currents in Pharmacy Teaching and Learning*, 13(11), 1451-1456.
<https://doi.org/https://doi.org/10.1016/j.cptl.2021.09.00>

Author Contributions

Paper 1: ATIC in SJTS

Nomi Reznik	Conceptualization, study designs, methodology, data collection, data analyses, writing, review and editing
Stefan Krumm	Conceptualization, study design, writing, review and editing, project supervision
Jan-Philipp Freudenstein	Conceptualization, study design (Study 1)
Anna-Luca Heimann	Conceptualization, review and editing
Pia Ingold	Conceptualization, review and editing
Philipp Schäpers	Conceptualization, study design (Study 1), review and editing
Martin Kleinmann	Conceptualization, study design (Studies 1 & 2), review and editing

Paper 2: SJTs and Their Components

Nomi Reznik	Conceptualization, study design, methodology, data collection, data analyses, writing, review and editing
Stefan Krumm	Conceptualization, study design, writing (review and editing), project supervision
Jan-Philipp Freudenstein	Conceptualization, study design
Philipp Schäpers	Conceptualization, study design

Paper 3: SJTs and Their Situation Construal

Nomi Reznik	Conceptualization, study design, methodology, data collection, data analyses, writing, review and editing
Stefan Krumm	Conceptualization, study design, writing (review and editing), project supervision
Alyce Thiel	Study Design, Methodology
Jan-Philipp Freudenstein	Conceptualization, study design

Eidesstattliche Erklärung

Hiermit erkläre ich,

- dass ich die vorliegende Dissertation selbstständig verfasst und ohne unerlaubte Hilfe angefertigt habe.
- dass ich die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen und KI-basierte Tools) entnommen sind, unter Angabe der Quelle kenntlich gemacht habe.
- Im Rahmen der Überarbeitung dieser Dissertation habe ich generative KI genutzt, um Inhalts- und Tabellenverzeichnisse zu erstellen und sprachliche Fehler (Rechtschreibung, Zeichensetzung) zu identifizieren. Nach der Nutzung dieser Tools habe ich den Inhalt überprüft und überarbeitet und übernehme die volle Verantwortung für den Inhalt der veröffentlichten Arbeit.
- Alle Hilfsmittel, die verwendet wurden, habe ich angegeben. Die Dissertation ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, den 08.09.2025

Nomi Reznik