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

Reznik, Nomi ; Krumm, Stefan ; Freudenstein, Jan-Philipp ; Heimann, Anna Luca ; Ingold, Pia ; Schäpers, Philipp ; Kleinmann, Martin

DOI: https://doi.org/10.1111/ijsa.12458

Originally published at:
Reznik, Nomi; Krumm, Stefan; Freudenstein, Jan-Philipp; Heimann, Anna Luca; Ingold, Pia; Schäpers, Philipp; Kleinmann, Martin (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.
DOI: https://doi.org/10.1111/ijsa.12458

Received: 12 October 2022 | Revised: 25 October 2023 | Accepted: 26 October 2023

DOI: 10.1111/ijsa.12458

RESEARCH ARTICLE

International Journal of Selection and Assessment | WILEY

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

Nomi Reznik[1] | Stefan Krumm[1] | Jan-Philipp Freudenstein[2] | Anna L. Heimann[3] | Pia Ingold[4] | Philipp Schäpers[5] | Martin Kleinmann[3]

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

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

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

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

[5]Department 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 "aaaaaaaaaa" 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 costruct 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.

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

ATIC scores			SJT item scores									
	M	*SD*	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 ATIC performance significantly predicted SJT performance, but with a small instead of the hypothesized moderate effect.

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

Predictors	Estimates	CI	*p*-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		
$\tau_{00\ Individual}$	0.03		
$\tau_{00\ SJT\ item}$	0.13		
$\tau_{11\ SJT\ item\ (ATIC)}$	0.00		
$\rho_{01\ SJT\ item}$	−0.04		
ICC	0.10		
Marginal R^2/conditional R^2	.001/.085		

Note: $N = 391$.

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

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

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 = *does not fit the construct at all* to 3 = *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

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											
	M	*SD*	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											
	M	*SD*	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 ($ICC_{clarity}$ = .59, κ_{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, κ > 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). *Ime4: 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., 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

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. (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*, 1–15. https://doi.org/10.1111/ijsa.12458