6-2020

# Removing situation descriptions from situational judgment test items: Does the impact differ for video-based versus text-based formats?

Philipp SCHÄPERS

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

Jan-Philipp FREUDENSTEIN
*Freie Universitat Berlin*

Joachim HÜFFMEIER
*Dortmund University*

Cornelius J. KÖNIG
*Universitat des Saarlandes*

# Removing Situation Descriptions From Situational Judgment Test Items: Does the Impact Differ for Video-Based Versus Text-Based Formats?

**Article** *in* Journal of Occupational and Organizational Psychology · October 2019

DOI: 10.1111/joop.12297

CITATIONS
0

READS
126

**6 authors**, including:

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

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

Jan-Philipp Freudenstein
Freie Universität Berlin
**9** PUBLICATIONS **1** CITATION
SEE PROFILE

Joachim Hüffmeier
Technische Universität Dortmund
**50** PUBLICATIONS **624** CITATIONS
SEE PROFILE

# Removing Situation Descriptions From Situational Judgment Test Items: Does the Impact Differ for Video-Based Versus Text-Based Formats?

Philipp Schäpers, Filip Lievens, Jan-Philipp Freudenstein, Joachim Hüffmeier, Cornelius J. König, & Stefan Krumm

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

This research was funded by the German Research Foundation (KR 3457/2-1). We acknowledge the help of Mareike Breda, Alexandra Göbel, Luca Kröger, Judith Pauly and Thomas Wilinski in collecting the data. We thank Barend Koch and Annemarie Goedbloed for providing the multimedia SJT materials and we acknowledge Ute-Christine Klehe for her valuable suggestions on an earlier version of this article.

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

This is the peer reviewed version of the following article: Schäpers, P., Lievens, F., Freudenstein, J.-P., Hüffmeier, J., König, C., & Krumm, S. (accepted). 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*, which has been published in final form at https://doi.org/ 10.1111/joop.12297, *Journal of Occupational and Organizational Psychology*. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

**Abstract**

Recent research has shown that many text-based situational judgment test (SJT) items can be solved even when the situational descriptions in the item stems are not presented to test takers. This finding challenges the traditional view of SJTs as low-fidelity simulations that rely on "situational" (context-dependent) judgment. However, media richness theory and construal level theory suggest that situation descriptions presented in a richer and more concrete format (video format) will reduce uncertainty about inherent requirements and facilitate the perception that the situation is taking place in the here and now. Therefore, we hypothesized that *situational* judgment would be more important in video situation descriptions than in text situation descriptions. We adapted a leadership SJT to realize a 3 (situation description in the item stem: video vs. text vs. none) × 2 (response format: video response options vs. text response options) between-subjects design ($N$ = 279). Participants were randomly assigned to one of the six conditions. The removal of video-based situation descriptions in item stems led to an equivalent decrease in SJT scores as the removal of text-based situation descriptions in item stems (video-based version: Cohen's $d$ = 0.535 vs. text-based version: Cohen's $d$ = 0.531). SJT scores were also contingent on the presentation format of both situation descriptions and response options: The highest scores were observed when situation descriptions and response options were presented in the same format. Implications for SJT theory and research are discussed.

*Keywords:* situational judgment test, contextualization, video, low-fidelity

**Practitioner Points**

- The presentation format did not moderate the effect of omitting situation descriptions in SJTs – i.e., the context-dependency of SJT performance did not increase when the SJT was administered in a video-based rather than a text-based format.
- The elimination of situation descriptions in item stems had a medium effect on overall test scores: SJT scores were significant lower without situation descriptions in comparison to SJT scores with situation descriptions (video-based version: Cohen's $d$ = 0.535 vs. text-based version: Cohen's $d$ = 0.531).
- It is important to match the stimulus and response formats in SJTs.

**Removing Situation Descriptions From Situational Judgment Test Items: Does the Impact Differ for Video-Based Versus Text-Based Formats?**

Situational Judgment Tests (SJTs) were reintroduced to the scientific community in the 1990s (Motowidlo, Dunnette, & Carter, 1990). Since then, they have become popular instruments for personnel selection and assessment. As their name suggests, SJTs have typically been portrayed as low-fidelity simulations that prompt situational judgments by requiring people to envision the presented job-related situations and judge how to respond to them.[1] However, recent findings have challenged this traditional view. Krumm et al. (2015) demonstrated that the majority of *text-based* SJT items could be solved even when the descriptions of job-related situation descriptions in the item stems were absent. Importantly, these findings call into question the "situational" nature of SJTs because they suggest that SJTs might operate more as measures of general (context-independent) domain knowledge than previously thought (Lievens & Motowidlo, 2016).

These findings have generated a heated scientific debate (e.g., Harvey, 2016; McDaniel, List, & Kepes, 2016; Melchers & Kleinmann, 2016; Naemi, Martin-Raugh, & Kell, 2016; Whetzel & Reeder, 2016). One of the conclusions from this debate was that Krumm et al. (2015) restricted their analysis to text-based SJTs, meaning that the definitive litmus test of the importance of situation descriptions in SJT item stems and thus of the context-dependency of SJT performance still had to be conducted. There are theoretical arguments for why *video-based* SJTs provide a better format for testing the importance of situation descriptions in SJT item stems. First, media richness theory (Daft & Lengel, 1986) suggests that richer media formats should be used in highly ambiguous situations—as is the case for SJTs—and that having situation descriptions in the item stems might make a

[1] In this paper, when we speak of responses to SJTs, we are referring to making a selection from among predetermined response options (closed answer format) and not to self-constructed responses (open answer format).

difference when richer media are used. Second and relatedly, construal level theory (Trope & Liberman, 2010) posits that abstract written information creates less situational immersion, leading to more general judgment and decision making. Thus, these two theories suggest that situation descriptions in SJT item stems may have been found to be less relevant in Krumm et al. due to their use of a suboptimal (text-based) stimulus format (and not because situation descriptions in item stems do not matter in SJTs *per se*). Therefore, Naemi et al. (2016) argued that video-based SJTs may be more "situational" than text-based SJTs and that "it is conceivable that this feature of video-based SJTs may allow the situational scenario composing the SJT item stem to have a greater impact on test takers' scores than traditional, text-based multiple response methods" (p. 79).

Hence, an important extension to previous research would be to explicitly consider the stimulus format (i.e., "the modality by which the test stimuli [e.g., information, questions, prompts] are presented to test-takers" Lievens & Sackett, 2017, pp. 45-46) when examining the relevance of situation descriptions in SJT item stems. In the current study, we did so by modifying not only the availability of situation descriptions in SJT item stems but also by modifying their stimulus format (video- vs. text-based). This study offers both theoretical and practical contributions. From a theoretical perspective, we contribute to a deeper understanding of a potential key boundary condition of SJTs' context-(in)dependency by investigating whether the results presented by Krumm et al. (2015) are valid only for situation descriptions at the lower end of the fidelity continuum (i.e., text-based SJTs). Moreover, this study is the first to test whether a key assumption underlying media richness theory (i.e., matching stimulus and response formats) make sense for SJTs. As a practical contribution, we are providing information to test developers about whether the cost-intensive development of video situations is worthwhile with regard to increasing the context-dependency of an SJT.

## Study Background and Hypothesis

### SJTs: Situations and Situational Judgment

SJTs are typically defined as low-fidelity simulations because they "present a verbal description of a hypothetical work situation, instead of a concrete representation, and … ask applicants to describe how they would deal with the situation, instead of having them actually carry out some action to deal with it" (Motowidlo et al., 1990, p. 640). It has been argued that SJTs function similarly to other simulations in that they are based on the behavioral consistency logic. That is, SJTs build on the notion of a point-to-point correspondence between simulated content and job requirements (Bruk-Lee, Drew, & Hawkes, 2013; Lievens & De Soete, 2012).

In line with the view of SJTs as simulations, the item stems—which present critical job-related situation descriptions that mirror pivotal aspects of the job—are considered to provide key content that is crucial for people's judgment processes when completing SJTs (Campion & Ployhart, 2013; Gessner & Klimoski, 2006). In fact, Weekley, Ployhart, and Holtz (2006) referred to the situation descriptions as the "bases for any SJT" (p. 158). Hence, many SJT guidelines provide detailed instructions for developing the situations in the item stems (e.g., McDaniel & Nguyen, 2001; Motowidlo et al., 1990; Weekley et al., 2006). In sum, situation descriptions in SJT item stems are typically regarded as an essential component of SJT items and for meaningful SJT responding.

Krumm et al. (2015) put this traditional view to the test in a series of studies, thereby investigating the impact of the situation descriptions in SJT item stems on SJT performance. They administered SJT items either with or without situation descriptions in the item stems. The absence of situation descriptions in SJT item stems should make it difficult for participants to apply context-dependent knowledge. Yet, Krumm et al. found that the presence of the situation description did not make a significant difference for many SJT items

(between 43% and 71% of the items). Furthermore, Krumm et al. were able to replicate these findings across different SJTs, response instructions (knowledge [should-do] vs. behavioral tendency [would-do]), and samples (of students and working people).[2] In sum, Krumm et al. concluded that the majority of SJT items are less context-dependent than previously assumed and that situation descriptions in the item stems may in fact not be as central to SJTs as typically thought.

Recently, these findings have generated quite a bit of controversy (Borneman, 2016; Brown, Jones, Serfass, & Sherman, 2016; Chen, Fan, Zeng, & Hack, 2016; Crook, 2016; Fan, Stuhlman, Chen, & Weng, 2016; Harris, Siedor, Fan, Listyg, & Carter, 2016; Harvey, 2016; Lievens & Motowidlo, 2016; McDaniel et al., 2016; Melchers & Kleinmann, 2016; Naemi et al., 2016; Torres & Beier, 2016; Whetzel & Reeder, 2016). One group of scholars echoed Krumm et al.'s (2015) call to reconceptualize SJTs as mainly context-independent measures (e.g., Crook, 2016; Harvey, 2016; Lievens & Motowidlo, 2016). They suggested that these recent findings justify the development of more generic and thus cost-effective SJTs that can be used across different job domains. Another group of scholars was more skeptical as to whether such far-reaching conclusions can be drawn from Krumm et al.'s results (e.g., Chen et al., 2016; Fan et al., 2016; Melchers & Kleinmann, 2016) because "more research is needed to determine the conditions under which situational scenarios are not required or

[2] Similar to the current study, Krumm et al. (2015) focused on situation descriptions (that are present in the item stems) instead of situations per se (present in both the item stems and response options). Because stripping off situation descriptions does not take out all the relevant context because the content of the response options might also be helpful for reconstructing the situation (see Melchers & Kleinmann, 2016). Krumm et al. also tested the influence of responses with context-specific versus context-independent response options. To illustrate, context-specific responses include context information that is related to the situation. An example of such a response is: "Declare an emergency, turn off all electrical systems, except for 1 NAVCOM and transponder, and continue to the regional airport as planned" which is a response to an aviation SJT (Hunter, 2003). Conversely, context-independent response options describe very general courses of action, such as: "I set specific and detailed goals" (Team Knowledge–KSA Test; Stevens & Campion, 1996). Importantly, Krumm et al. found only mixed evidence that the content of the response options moderated the results for SJT items that assessed applied social skills (i.e., the construct domain including leadership skills; see Christian, Edwards, & Bradley, 2010). In one of two studies, performance in SJTs addressing applied social skills did not differ for response options denoting context-specific courses of action compared with general courses of action when situation descriptions were omitted.

necessary" (McDaniel et al., 2016, p. 49). For example, Naemi et al. raised the question of whether Krumm et al.'s findings applied equally to (more realistic) video-based SJTs. To answer this question in the current study, we examined whether the importance of situation descriptions for SJT performance is moderated by the SJT's stimulus format.

**Why Should Situation Descriptions Matter More in Video-Based SJTs?**

There are at least two reasons why situational descriptions might matter more when completing video-based SJTs. One reason is that situation descriptions presented in a video format more closely resemble the real world than text-based situations (MacCann, Lievens, Libbrecht, & Roberts, 2016; Naemi et al., 2016). For example, Olson-Buchanan and Drasgow (2006) emphasized that video-based SJT formats "provide a much richer assessment environment that allows the situational context to be richly portrayed" (p. 253).

Conceptually, this first reason fits well with the rationale behind media richness theory (Daft & Lengel, 1984; Fulk & Boyd, 1991) in which ambiguity serves as a central concept. Communication media are ordered along a continuum of media richness on the basis of their capacity to transmit information and resolve this ambiguity. In particular, media richness theory posits that richer media can be distinguished from leaner media on the basis of four specific factors: opportunity for two-way communication (feedback), ability to convey a multiplicity of cues (verbal and nonverbal), ability to convey a sense of personal focus, and use of natural language. Essentially, these factors refer to the medium's ability to carry a variety of data (e.g., aural cues, visual cues, text cues) and to carry symbolic information (e.g., emotions) from and about the individuals who are communicating. The basic premise of media richness theory is that communication media are most efficient when they match the degree of ambiguity present in the task and situation. In other words, the medium should fit the type of message. Richer media (e.g., face-to-face, video) should be

used when ambiguity is high, whereas leaner media (e.g., text) are sufficient when ambiguity is low.

It seems doubtful that a written medium would be able to match the level of ambiguity inherent in SJT items because a written medium cannot convey the various cues that are present in social interactions (e.g., body language, tone of voice, and inflection; cf. McDaniel et al., 2016). Conversely, in video-based SJTs, test takers are provided with verbal/nonverbal/paralingual cues (e.g., information about body language, facial expressions, intonation, and pitch of voice) and emotional cues that are typically not present in text-based SJTs. When video-based situation descriptions were not present, meaning that this wealth of information was no longer available to test takers, we expected scores on the SJT to be lower than when video-based situation descriptions were presented.

A second and related reason is that video situation descriptions should lead to lower psychological distance perceptions among test takers. This assumption is rooted in construal level theory (Trope & Liberman, 2010), which posits that objects and situations may be construed on a continuum ranging from abstract (high-level construal) to concrete (low-level construal). For instance, referring to “a co-worker” in a text-based situation description is more abstract than presenting a specific person in a video because the latter conveys information about, for example, age, gender, and height. A key assumption of construal level theory is that abstract, high-level construals “bring to mind more distal instantiations of objects. For example, ‘having fun,’ compared with ‘playing basketball outside,’ may bring to mind activities in the more distant future and past, in more remote locations, in hypothetical situations, and with more socially distant others” (Trope & Liberman, 2010, p. 442). In other words, the more abstract the presentation of a situation in an SJT, the more test takers might rely on their general past experiences and general preferences. Furthermore, they might be

less inclined to envision themselves in a specific situation and judge how they would act if this situation were happening in the here and now.

Initial evidence for the greater importance of video situation descriptions was provided by Rockstuhl, Ang, Ng, Lievens, and Van Dyne (2015). These authors expanded the traditional SJT paradigm not only by administering the typical SJT response effectiveness instructions ("what would/should you do in the given situation") but also by asking how participants actually perceived and interpreted the video situations. They found that "understanding the intentions, emotions, and thoughts of the parties in the situation were the dominant types of situational judgments" (p. 475). In addition, test takers' construal of the video situation descriptions predicted traditionally derived SJT scores and provided incremental predictive validity above and beyond judgments of response effectiveness. Although they did not explicitly compare written to video-based situation descriptions, the authors' results underlined the importance of providing and judging video situation descriptions. In light of the above theoretical considerations, we expected that the absence of situation descriptions in a video-based SJT would lead to a larger decrease in SJT scores than the absence of situation descriptions in a text-based SJT.

Note, however, that the greater performance decline expected in video-based versus text-based SJT scores (due to the absence of the situation description) may be masked by the fact that video-based SJTs may also present *response options* in a video-based format, which might provide additional information in a richer format than text-based SJTs. Hence, the absence of situation descriptions in video-based SJTs may result in only small decreases in SJT scores because there is still a lot of contextual information included in the video response options (see Harris et al., 2016; Kaminski, Felfe, Schäpers, & Krumm, 2019; Leeds, 2012; Melchers & Kleinmann, 2016). Therefore, the relevance of situation descriptions in video- versus text-based SJTs cannot be determined without considering the format of the response

options. To examine whether potential differences between video- and text-based SJTs may be co-determined by the response option format, we decided to manipulate not only the presentation of situation descriptions (video situation vs. text situation vs. no situation) but also the response format (video responses vs. written responses).

On the basis of the conceptual and empirical arguments presented above, we believe that video-based SJTs might "present a case in which the situational content of SJTs matters, as these SJT formats may compensate for the construct underrepresentation (Messick, 1995) of text-based SJTs by measuring test takers' ability to accurately perceive situations" (Naemi et al., 2016, p. 81). Hence, we posited:

> *Hypothesis 1 (H1): The absence of situation descriptions in a video-based SJT will lead to a larger decrease in SJT scores than the absence of situation descriptions in a text-based SJT.*

## Method

### Sample

Participants were recruited via online postings (on Facebook, university websites, and in newsletters), email, and poster advertising in a large German city. Inclusion criteria were: age 18 or older, a minimum level of leadership experience, and English language ability equal to or higher than *Level B1* according to the Common European Framework of Reference for Languages (because the study was administered in English). Individuals interested in participating in the current study first had to complete an English language test (University of Cambridge Local Examinations Syndicate, 2016), which was administered online. A score of at least 16 correct answers out of 25 possible (equivalent to Level B1) was required for participation,[3] which resulted in the exclusion of 78 individuals. In addition,

[3] The results reported below did not change when we controlled for English language ability. The reliability of the English language test scores was $\alpha = .82$.

potential participants had to provide information about their leadership position and duration of leadership experience. Another 49 individuals were not eligible for participation because they reported having no leadership experience (or did not respond to the assessment of leadership experience).[4]

The final sample comprised 279 participants (74.2% female) with a mean age of 26.19 years (*SD* = 7.44, range 18 to 66).[5] Among these, 92% had at least six months of leadership experience (average leadership experience = 4.62 years, *SD* = 5.05), and 69.9% reported a moderate or higher degree of leadership experience on a 6-point Likert scale. Regarding education levels, 56.3% of the participants held a university entry degree (comparable to A-levels), and 35% held a university degree. Participation was compensated with 15€ or credits for university students majoring in psychology (28% of participants). Voluntary participation and anonymity were ensured.

**Study Design and Materials**

In our quasi-experimental study, we used a 3 (situation description in the item stem: video vs. text vs. none) × 2 (response format: video response options vs. written response options) between-subjects design to test H1. Participants were randomly assigned to one of the six conditions, which differed solely in the SJT version that was administered.[6] As much as possible, everything else was kept constant. Extraneous variables were controlled for, as much as possible, through the laboratory setting and the randomized allocation of participants to conditions. The distribution of participants across the six conditions was approximately equal (between 41 to 50 participants per cell). The assessment was conducted in proctored

[4] Leadership experience was an inclusion criterion for study participation to ensure that the leadership SJT was meaningful for participants (see study design and materials below).

[5] An a priori power analysis (G*Power; Faul, Erdfelder, Lang, & Buchner, 2007) revealed that 251 participants were necessary to test the hypothesis with sufficient power ($1 - \beta = .95$; $\alpha = .05$). On the basis of Krumm et al.'s (2015) results, we assumed a moderate effect size of $f = .25$. Using an $F$ test for ANOVA fixed effects, special, main effects, and interactions, G*Power returned $\lambda = 16.688$ and a critical $F$-value of $F(2, 245) = 3.032$.

[6] Randomization of the test conditions was realized by randomizing the test sessions. Specifically, the authors of this manuscript used a computer-generated chance algorithm to make an a priori assignment of each session to one of the six conditions. Thus, all participants in each test session worked on the same test condition.

group sessions (up to nine individuals were tested at the same time) at a comprehensive state university in Germany. In addition to the SJT, participants also completed a test-taking motivation scale. All tests were administered in English.

**Situational Judgment Test.** We used an SJT in the leadership domain originally developed in a video-based format by Oostrom, Born, Serlie, and van der Molen (2012). It consisted of 17 short videotaped vignettes of key interpersonal situations that managers might face in their job (e.g., developing teams, coordinating and motivating employees, decision making, negotiating skills, and conflict management; Peterson, Borman, Mumford, Jeanneret, & Fleishman, 1999). After watching the scenarios, participants were asked to evaluate the effectiveness of each of four possible reactions (also presented in a video format in the original version). A sample item is presented in Appendix A. Participants rated the effectiveness of each response on a 5-point scale ranging from – – = *very ineffective* to + + = *very effective.* We used the expert scoring key developed by the test authors (Oostrom et al., 2012). That is, we calculated each participant's absolute deviation from the expert rating. As recommended by the test authors, we used the aggregated absolute deviation (across all responses) as the dependent variable. It took participants about 45 minutes to complete this SJT (regardless of experimental condition).

In addition to the original version of this SJT, which presented both the situation descriptions and response options in a video format, five additional versions were created to operationalize all of the cells in the 3 × 2 quasi-experimental design. Text versions of the situation descriptions in item stems and the response options were created by transcribing the original video versions. Note that nonverbal behavior was not described in the text versions of the SJT (as suggested by Lievens & Sackett, 2006). Stemless versions of the SJT were created by omitting the situation descriptions in the item stems (in line with Krumm et al., 2015). Hence, these versions included only response options in either a video or text format.

**Test-taking motivation.** SJT items without item stems (i.e., situation descriptions) might represent an unexpected format for participants. In addition, participants received less information to guide their response choice in these conditions. This might cause frustration and confusion, which may in turn potentially lead to lower test-taking motivation. To check whether test-taking motivation differed across conditions, every participant completed five items from the Test Attitude Survey (TAS; Arvey, Strickland, Drauden, & Martin, 1990) at the end of the survey. A sample item is: "I was extremely motivated to do well on this test or tests." Participants responded on a 5-point Likert scale from *disagree strongly* (1) to *agree strongly* (5). The internal consistency of this scale was acceptable ($\alpha = .78$).

### Data Analyses

A one-way analysis of variance (ANOVA) was used to test whether the absence of situation descriptions in the item stems in a video-based SJT led to a larger decrease in SJT scores than the absence of situation descriptions in the item stems in a text-based SJT. We used a two-way ANOVA with a subsequent linear contrast analysis to test whether potential differences between the video and text versions were moderated by the modality of the response format. For analyses on the item level, we conducted an independent samples *t* test per item to compare SJT performance for items with and without situation descriptions. Following established recommendations, we used eta-squared ($\eta^2$) as the effect size for the ANOVAs. Cohen's *d* and $\eta^2$ were reported for potential differences on the item level (e.g., Cohen, 1973, 1988; Pierce, Block, & Aguinis, 2004). Unless otherwise described, the data were analyzed using SPSS (version 24).

## Results

### Preliminary Analyses

We began by checking whether the six groups differed on demographic, psychological, or skill-based variables of interest. The six groups did not differ significantly

in terms of age, $F(5, 273) = .280$, $p = .92$, $\eta^2 = .005$, gender, $\chi^2(5) = 1.599$, $p = .90$, $\phi = .08$, level of education, $\chi^2(35) = 32.610$, $p = .58$, $\phi = .32$, years of leadership experience, $F(5, 271) = .569$, $p = .72$, $\eta^2 = .010$, personality facets (for a description of the measure, see Rammstedt & John, 2005), $F(25, 1365) = .824$, $p = .71$, $\eta^2 = .015$, or English language skills, $F(5, 273) = 1.163$, $p = .33$, $\eta^2 = .021$. Test-taking motivation also did not differ across conditions, $F(5, 273) = .567$, $p = .73$, $\eta^2 = .010$.

Next, we inspected the reliabilities of the SJT scores in each condition. Cronbach's α and McDonald's ω showed good or acceptable estimates for SJT scores in all conditions. The reliability of the SJT scores in the six conditions ranged from .63 to .87 (Cronbach's α) and from .58 to .87 (McDonald's ω total[7]). These reliability estimates are above those reported in meta-analyses on SJTs in general (Catano, Brochu, & Lamerson, 2012; Kasten & Freund, 2016; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). SJT scores for one condition (video situation description/text responses) exhibited a lower internal consistency ($\alpha = .63$, $\omega = .58$) than the others. According to the test for differences between alphas (Feldt, Woodruff, & Salih, 1987; R-package cocron, version 1.0.-1; Diedenhofen & Musch, 2016), the reliabilities of the SJT scores for the remaining conditions did not differ significantly from each other, $\chi^2(4) = 6.028$, $p = .20$.

Finally, to ensure that the measured SJT scores elicited similar response patterns across the six SJT conditions with and without situation descriptions, we conducted multiple-group measurement invariance analyses by applying maximum-likelihood estimation using R (version 3.4.0; R Core Team, 2017) and the R-package lavaan (version 0.5–22; Rosseel, 2012). In line with the measurement model of the SJT as established by the test authors (Oostrom et al., 2012), we specified a one-factor model and tested for metric invariance (i.e.,

[7] McDonald's ω total was calculated using R (version 3.4.0; R Core Team, 2017) and the R package userfriendlyscience (version 0.7.1; Peters, 2015).

invariant factor loadings across groups).[8] In light of model complexity and our sample size (see MacCallum, Widaman, Zhang, & Hong, 1999; Worthington & Whittaker, 2006), we used an item-parceling procedure to ensure model identification. In accordance with Little, Cunningham, Shahar, and Widaman's (2002) recommendations, the 17 SJT items were randomly divided into six item parcels. Therefore, we specified a one-factor model with one latent factor (overall SJT performance) and six indicator variables (one indicator represented one of the six parcels). We used common fit criteria to evaluate the model fit (Beauducel & Wittmann, 2005; Browne & Cudeck 1993; Byrne, 1989; Hu & Bentler, 1999, Kline, 2004). We considered model fit to be acceptable for the following values: comparative fit index (CFI) > .90, root mean square error of approximation (RMSEA) < .10 (preferably < .05), and standardized root mean square residual (SRMR) < .10. A $\chi^2$-difference test was used to evaluate the relative fit of the nested models.

The baseline model (no restrictions) showed a good fit, with $\chi^2(54) = 78.442$, CFI = .943, RMSEA = .099, and SRMR = .061. In addition, when we tested for metric invariance (all factor loadings restricted), the model fit did not decrease substantially, $\Delta\chi^2(25) = 32.500$, $p = .14$, CFI = .925 RMSEA = .093, and SRMR = .106.[9] Therefore, metric invariance could be assumed across the different conditions, meaning that the measurement structure was invariant across the six conditions. This was a necessary prerequisite for interpreting the following between-group differences (Bollen, 1989).

**Hypothesis Tests**

[8] We also specified a four-factor model (with the following factors: addressing results, addressing social behavior, motivating, coaching). However, the results obtained for the original SJT version (video situation descriptions and video responses) revealed a non-positive definite covariance matrix of latent variables and a poor model fit (CFI = .78, RMSEA = .084, SRMR = .104).

[9] Although the measurement of constructs in SJTs is sometimes referred to as a "hot mess" (see McDaniel et al., 2016), some notable exceptions have yielded measurement models that have been well-aligned with their theoretical structure (e.g., Bledow & Frese, 2009; Gatzka & Volmer, 2017; Mussel, Gatzka, & Hewig, 2018; see also Guenole, Chernyshenko, & Weekly, 2017).

We hypothesized that the absence of situation descriptions in the item stems in video-based SJTs would lead to a stronger decrease in SJT scores than the absence of situation descriptions in the item stems in text-based SJTs (H1). As an overall test, we conducted a two-way ANOVA to test whether there was an interaction between the SJT version (video- vs. text-based SJT) and the presence of situation descriptions in the item stems (with vs. without situation descriptions in the item stems). Because H1 addressed text- versus video-based SJTs, this analysis focused on only the experimental conditions that included SJTs with congruent modalities (e.g., video situations and video responses). In other words, we focused on "pure" text-based and "pure" video-based SJTs. This was also done because video-based SJTs often consist of video situation descriptions in item stems and video response options, whereas text-based SJTs typically comprise text situation descriptions in item stems and text response options. We thereby analyzed only a 2 (SJT modality: video SJT vs. text SJT) × 2 (situation description: with situation description vs. without situation description in item stem) version of our design in this analysis.

Results showed a significant main effect for SJT version: Participants who completed the video-based SJT version obtained a higher SJT score than those who worked on the text-based SJT, $F(1, 180) = 5.841$, $p = .017$, $\eta^2 = .029$. Furthermore, there was a main effect for situation descriptions in item stems (i.e., the absence of situation descriptions in item stems led to a lower SJT score), $F(1, 180) = 12.999$, $p < .001$, $\eta^2 = .065$, indicating that situation descriptions matter for SJT performance. Importantly for H1, there was no significant interaction between SJT version and situation descriptions in the item stems, $F(1, 180) < 0.001$, $p = .990$, $\eta^2 < .001$, suggesting that the absence of situation descriptions in item stems did not differently affect the performance in video- and text-based SJTs (see Figure 1). Specifically, omitting situation descriptions in the video-based SJT resulted in an effect size

of Cohen's $d = 0.535$; omitting situation descriptions in the text-based SJT had a similar effect (Cohen's $d = 0.531$).

To rule out the possibility that the response option format (video vs. text) influenced the above results, we also conducted an overall test on the fully crossed 3 (situation description: video situation description vs. text situation description vs. no situation description) × 2 (response format: video responses vs. written responses) design. We again found a main effect for situation descriptions in item stems, $F(2, 273) = 7.346$, $p = .001$, $\eta^2 = .049$, and for response format, $F(1, 273) = 12.184$, $p = .001$ $\eta^2 = .041$. Following the conventions for interpreting effect sizes ($\eta^2$ cut-off values: small $\eta^2 < .06$, medium $.06 \leq \eta^2 < .14$, large $\eta^2 \geq .14$; for further information, see Cohen, 1973, 1988), our findings can be considered to represent small to medium effects. More importantly, there was no significant interaction between situation descriptions in item stems and response format, $F(2, 273) = 0.693$, $p = .501$, $\eta^2 = .001$.

Finally, we examined the effect of omitting video- and text-based situation descriptions in the item stems while keeping the response format constant. In a first analysis, we included only SJT versions with video-based response options. We assumed a linear decrease in SJT performance (i.e., video situation descriptions > text situation descriptions > no situation descriptions). To test for this linear trend, we conducted a one-way ANOVA with a subsequent linear contrast analysis. The three SJT scores differed significantly, $F(2, 132) = 3.957$, $p < .05$, $\eta^2 = .057$. We found a linear contrast (video SJT > written SJT > no situation descriptions), $t(132) = 2.700$, $p < .05$, which supported H1.

In a second analysis, we scrutinized only SJT versions with text-based response options. As homogeneity of variances was violated, $F(2, 141) = 4.797$, $p < .05$, we used Welch's $F$. Again, the ANOVA produced a significant main effect for situation description (video situation, text situation, no situation), Welch's $F(2, 88.92) = 3.225$, $p < .05$, $\eta^2 = .055$.

However, contrast analyses did not support the assumed linear trend (video situation descriptions [1], text situation descriptions [0], no situation descriptions [-1]), $t(71.66) = 1.704$, $p = .093$. Rather, results suggested a different linear trend: written SJT scores > video SJT scores > no situation descriptions, $t(80.86) = 2.546$, $p < .05$. Considering this along with the results obtained for video-based response options, the congruency between an SJT's situation description and response modality format (instead of the format per se) seemed to be an important determinant of SJT performance (see Figure 2).

**Ancillary Analyses**

Apart from our main analyses, we also inspected differences between the SJT with or without situation descriptions in item stems at the item level (see Krumm et al., 2015) and did this separately for the video- and text-based SJT version. Results revealed that 8 out of 17 video-based SJT items and 9 out of 17 text-based SJT items yielded significantly lower scores when administered without situation descriptions in the item stems.[10] This indicates that at the item level, it did not make a significant difference whether situation descriptions were present or absent for 47% to 53% of the items. We used the very liberal approach with an unadjusted alpha level (beyond the approach with an adjusted alpha level) to account for the lower reliability of item scores in comparison with overall scores, which may otherwise mask potential differences between items with and without situation descriptions.[11] When the alpha level was adjusted to account for multiple significance tests,[12] it did not make a significant difference for 15 of the 17 video-based items (i.e., for 88%) whether situational descriptions were included in the item stems or not. The same result was obtained for text-based items (see Tables 1 and 2). The average effect size across items was $\eta^2 = .04$ (video-

[10] Among these 8 and 9 items, 5 items were identical across the video- and the text-based SJT versions.
[11] Results did not differ when we used an even more liberal approach with $p < .10$ (as suggested by an anonymous reviewer).
[12] We used the Bonferroni correction (Cabin & Mitchell, 2000) and divided the *p*-value by the number of tests (.05/17 = .00294).

based; range .00 to .11) and $\eta^2 = .03$ (text-based; range .03 to .16). In sum, the item-level results provided further evidence that most SJT items can be "placed on a continuum with some SJTs measuring rather context-independent knowledge and others being situated on the context-dependent knowledge side" (Krumm et al., 2015, p. 404).

Furthermore, we also investigated zero-order correlations of the SJT scores in the six conditions with ratings on broad personality dimensions (Rammstedt & John, 2005) and emotional intelligence—including the three subtests: emotion perception, emotion understanding, and emotion regulation/management (Allen et al., 2015; Allen, Weissman, Hellwig, MacCann, & Roberts, 2014; Schlegel, Grandjean, & Scherer, 2014; Schlegel & Scherer, 2016). We did so to detect potential differences in construct saturation in SJT scores across the conditions. For all comparisons of correlation coefficients, we applied Fisher's *z* transformation (Cohen, Cohen, West, & Aiken, 2003) so that the difference in the respective *z* scores could be tested for statistical significance. Interestingly, text-based SJT versions exhibited higher correlations with emotional intelligence than the video-based SJT version. Yet, no significant differences in correlations occurred when video- or text-based situation descriptions in the item stems were omitted, neither for emotional intelligence ($z$s = |.22| to |1.02|, $p$s = .15 to .41 for the video-based SJT versions; $z$s = |.30| to |1.20|, $p$s = .12 to .38 for the text-based SJT version) nor for personality ($z$s = |.05| to |1.53|, $p$s = .06 to .29 for the video-based SJT versions; $z$s = |.09| to |1.20|, $p$s = .12 to .47 for the text-based SJT version; see Appendix B).

Finally, we also followed up on an anonymous reviewer's suggestion and explored whether test takers would be able to reconstruct the content of the situation description only on basis of the response options (see also Melchers & Kleinmann, 2016). Thus, we presented only the response options along with six to eight statements about what the situation description might have consisted of. Note that we made sure that 50% of the statements

represented correct situational information (i.e., information that was part of the actual situation description) and the other 50% incorrect information (i.e., information that was not part of the situation description). We asked eight raters (50% female with a mean age of 27.13 years; *SD* = 3.23) to indicate whether they thought that the statements represented parts of the situation or not.

The results of this signal detection task revealed mixed evidence. The percentage of correctly assigned statements varied from 53% to 91% per item. Thus, it seemed that participants could sometimes reproduce (large) parts of the context based on only the responses. However, importantly, we found no significant relation between the percentage of correctly identified situational content and differences in performance between SJTs with versus without situation descriptions. To examine this, we used the mean percentage of correctly identified situational content per item as a new variable and correlated this variable with the effect sizes given in Table 2 ($r$ = .161; i.e., differences in SJT performance with vs. without situation descriptions). Furthermore, we found that the percentage of correctly identified situational content was also not related to SJT performance in general. For this analysis, we compared the mean performance on the item level of the SJT version without situation descriptions with the mean percentage of correctly identified situational content across the eight raters. There was no significant correlation ($r$ = -.142)[13]. Thus, response options seem to enable test takers to reconstruct contextual information related to the situation to some extent. However, this reconstruction of contextual information on the basis of response options does not seem to be systematically associated with SJT performance.

## Discussion

In this study, we examined whether the absence of situation descriptions in a video-based SJT would lead to a larger decrease in SJT scores than the absence of situation

[13] Please note that reported correlations are only based on a small sample size.

descriptions in a text-based SJT. Contrary to H1, we did not find a significant interaction between SJT format (video- vs. text-based) and the availability of situation descriptions in item stems (i.e., the scenarios), even when we controlled for the modality of response options. In other words, the absence of video situation descriptions in item stems and the absence of text situation descriptions in item stems resulted in a similar decrease in SJT performance (video-based version: Cohen's $d$ = 0.535 vs. text-based version: Cohen's $d$ = 0.531).

**Implications for Theory**

Our findings have several implications for SJT research and theory. First, this study provides insights into how test takers solve video-based SJTs. Given that multilayered cues (e.g., tone of voice, body language, facial expressions) were available in the situation descriptions for participants completing the "full" video version of the SJT, one might have assumed that omitting these cues would lead to a large decrease in SJT performance. However, omitting video situation descriptions did not lead to a greater decrease in performance than omitting text situation descriptions. In an examination of a scenario-based social intelligence test, Baumgarten, Süß, and Weis (2015) reported somewhat similar findings: They showed that contextual information (e.g., age and gender of the acting person) did not improve participants' performance. In a similar vein, Gesn and Ickes (1999) revealed that judgments about other people's thoughts and feelings were equally accurate when based on only audio material compared with both video and audio recordings. These authors explained their findings on the basis of significant clue theory (Archer & Akert, 1980), which posits that in certain contexts, some cues (in this case auditory cues) may carry the most meaning and are thus mostly used to make judgments. In the case of the video SJT used in this study, test takers may have found that the most meaning was conveyed by the actual dialogues (which were the same in the text-based and video-based SJTs). In other words, they

may have perceived the additional non-verbal content provided by the video format as less diagnostically useful.

Second, our results offer insights about the interaction between the modalities of response options and item stems. According to media richness theory (Daft & Lengel, 1986; Potosky, 2008), "performance will be improved when task information needs are matched to a medium's information richness" (Dennis, Fuller & Valacich, 2008, p. 575). Thus, media richness theory posits that SJT performance will be better when both the response options and situation descriptions are presented either as video clips or texts. In this respect at least, our findings are in line with media richness theory and suggest that the congruence between the modalities of the response options and the situation descriptions affects SJT performance. That is, average scores on SJT versions in which the response options and situation descriptions matched (i.e., either both presented in a video format or both in a written format) were higher than scores on an incongruent SJT version in which video situation descriptions were for instance combined with textual response options.

**Implications for Practice**

As a first practical implication, we advise test developers to consider whether the expected benefits of video SJTs justify the costs associated with their development. The well-documented advantages of video SJTs (e.g., higher face validity, Chan & Schmitt, 1997; improved candidate involvement, Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000; higher validity for predicting interpersonal criteria, Christian et al., 2010) must be weighed against the finding that video and text-based SJTs did not differ in terms of the importance of situation descriptions. In general, we would like to emphasize that we do not recommend using SJTs without situations and are not positing that SJTs without situation descriptions are the panacea.

A second practical implication is related to whether it makes sense to combine different stimulus and response format modalities. Many SJTs are indeed "pure" versions, consisting entirely of either text or video material. Although test developers might be tempted to reduce costs by creating hybrid SJT versions, our findings indicate that a mixture of modalities in SJTs (i.e., video situation descriptions alongside text responses and vice versa) is not recommendable. Instead, it is best to keep the axiom of media richness theory into account and use the same stimulus and response format (see also Lievens & Sackett, 2017).

A final practical implication pertains to the methodological approach used in this study. In line with recent calls for alternative test validation strategies (beyond inspecting correlation matrices; e.g., Bornstein, 2011; Borsboom, Mellenbergh, & van Heerden, 2004), we adopted an experimental validation procedure. That is, we manipulated several key features of a test and examined whether this manipulation affected test performance (see Krumm, Hüffmeier, & Lievens, 2017, for more details). This procedure revealed valuable insights into the inner workings of video SJTs and should thus be more frequently adopted by test developers. Such procedures can also be regarded as a kind of "manipulation check" and might be applied to select video scenarios that actually are crucial for SJT performance.

**Limitations and Future Research**

Several limitations of the present study should be acknowledged. First, the experimental conditions selected differ in several ways. That is, the condition without situation descriptions in the item stems might have done more to participants than simply exclude situation descriptions for them. For instance, the lack of information may have demotivated participants (see also Chan, Schmitt, DeShon, Clause, & Delbridge, 1997). That is, test takers might have become increasingly frustrated by the difficulty of reconstructing the situation description due to the lack of information inherent in SJT items without situation descriptions. This might have in turn reduced their test-taking motivation when completing

the SJT items. However, we did not find differences in test motivation across groups. Second, our findings are based only on one SJT dealing with a single construct domain (leadership). Hence, one might question whether our results are transferable to other SJTs. Third, this study was administered in a low-stakes context. Thus, one may speculate about its generalizability to a high-stakes testing situation (but see Attali, 2016).

In terms of future research, we encourage similar research with other SJT stimulus formats (e.g., 3D animated, virtual reality, avatar-based). For instance, virtual-reality SJTs create a strong feeling of presence in the situation and allow test takers to interact with the situation (North, North, & Coble, 2002). Relatedly, future research should shed light on when and why more realistic presentation formats contribute to SJT performance and validity. In fact, we do not know how important video situation descriptions are for the predictive potential of SJT scores, their relations with other constructs, subgroup differences, or applicant perceptions. Such research might also show which formats create a sense of involvement in the presented situation (see construal level theory; Trope & Liberman, 2010). First evidence revealed that construct saturation, applicant perceptions, and the prediction of global criteria were only little affected by removing situation descriptions in text-based SJTs (Schäpers et al., 2019). Additional research would be useful to clarify if that also applies to video-based SJTs. Future research might further clarify whether the content of response options enables to construe the missing situation description. While we found initial evidence that the response options are also valid sources of situational information using a small sample of raters, more comprehensive tests of this hypothesis are needed. Currently, it is still unclear which features of response options are crucial for meaningful SJT responding and whether specific features (e.g., degree of contextualization, response format, response modality; see also Lievens & Sackett, 2017) can compensate for the absence of situation descriptions in the item stems.

## Conclusion

This study contributed to the current debate about the conceptualization of SJTs as context-(in)dependent selection procedures. We found that the removal of video- or text-based situation descriptions in item stems had a medium effect on overall SJT scores. Notably, our study also revealed that the removal of video-based situation descriptions in item stems led to an equivalent decrease in SJT scores as the removal of text-based situation descriptions in item stems. Additionally, we found evidence for the importance of matching stimulus and response formats in SJTs.

## References

Allen, V., Rahman, N., Weissman, A., MacCann, C., Lewis, C., & Roberts, R. D. (2015). The Situational Test of Emotional Management–Brief (STEM-B): Development and validation using item response theory and latent class analysis. *Personality and Individual Differences, 81,* 195-200. https://doi.org/10.1016/j.paid.2015.01.053

Allen, V. D., Weissman, A., Hellwig, S., MacCann, C., & Roberts, R. D. (2014). Development of the Situational Test of Emotional Understanding–Brief (STEU-B) using item response theory. *Personality and Individual Differences, 65,* 3-7. https://doi.org/10.1016/j.paid.2014.01.051

Archer, D., & Akert, R. M. (1980). The encoding of meaning: A test of three theories of social interaction. *Sociological Inquiry, 50,* 393-419. https://doi.org/10.1111/j.1475-682X.1980.tb00028.x

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

Attali, Y. (2016). Effort in low-stakes assessments: What does it take to perform as well as in a high-stakes setting? *Educational and Psychological Measurement, 76,* 1045-1058. https://doi.org/10.1177/0013164416634789

Baumgarten, M., Süß, H. M., & Weis, S. (2015). The cue is the key: The relevance of cues and contextual information in the social understanding tasks of the Magdeburg Test of Social Intelligence. *European Journal of Psychological Assessment, 31,* 38-44. https://doi.org/10.1027/1015-5759/a000204

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

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

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

Borneman, M. J. (2016). Further considerations in SJT development. *Industrial and Organizational Psychology, 9*, 55-59. https://doi.org/10.1017/iop.2015.117

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

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071. https://doi.org/10.1037/0033-295X.111.4.1061

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

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

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

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

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

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

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

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

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300-310. https://doi.org/10.1037/0021-9010.82.2.300

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

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

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33*, 107-112. https://doi.org/10.1177/001316447303300111

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9780203771587

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

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

Daft, R. L., & Lengel, R. H. (1984). Information richness: A new approach to manage information processing and organizational design. In L.L. Cummings & B.M. Staw (Eds.), *Research on organizational behavior* (Vol. 6, pp. 191–233). Greenwich, CT: JAI Press.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32,* 554-571. https://doi.org/10.1287/mnsc.32.5.554

Dennis, A. R., Fuller, R. M., & Valacich, J. S. (2008). Media, tasks, and communication processes: A theory of media synchronicity. *MIS Quarterly, 32*, 575-600. https://doi.org/10.2307/25148857

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science, 11,* 51-60.

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

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

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11,* 93-103. https://doi.org/10.1177/014662168701100107

Fulk, J., & Boyd, B. (1991). Emerging theories of communication in organizations. *Journal of Management, 17,* 407-446. https://doi.org/10.1177/014920639101700207

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

Gesn, P. R., & Ickes, W. (1999). The development of meaning contexts for empathic accuracy: Channel and sequence effects. *Journal of Personality and Social Psychology, 77,* 746-761. https://doi.org/10.1037/0022-3514.77.4.746

Gessner, T. L., & Klimoski, R. J. (2006). Making sense of situations. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 13-83). Mahwah, NJ: Erlbaum.

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

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

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

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

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *International Journal of Aviation Psychology, 13,* 373–386. https://doi.org/10.1207/S15327108IJAP1304_03

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,* 72–82. https://doi.org/10.1111/ijsa.12233

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

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

Krumm, S., Hüffmeier, J., & Lievens, F. (2017). Experimental test validation: Examining the path from test elements to test performance. *European Journal of Psychological Assessment,* e-pub ahead of print. https://doi.org/10.1027/1015-5759/a000393

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

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

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

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9,* 3-22. 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,* 1181-1188. https://doi.org/10.1037/0021-9010.91.5.1181

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

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

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84-99. https://doi.org/10.1037/1082-989X.4.1.84

MacCann, C., Lievens, F., Libbrecht, N., & Roberts, R. D. (2016). Differences between multimedia and text-based assessments of emotion management: An exploration with the multimedia emotion management assessment (MEMA). *Cognition and Emotion, 30,* 1317-1331. https://doi.org/10.1080/02699931.2015.1061482

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

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

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

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741

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

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

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

North, M. M., North, S. M., & Coble, J. R. (2002). Virtual reality therapy: An effective treatment for psychological disorders. In: K. Stanney (Ed.), *Handbook of virtual environments* (pp. 1065–1079). New York, NY: Erlbaum.

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

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

Peters, G. J. Y. (2015). userfriendlyscience: Quantitative analysis made accessible. R package version 0.3-0.

Peterson, N. G., Borman, W. C., Mumford, M. D., Jeanneret, P. R., & Fleishman, E. A. (1999). *An occupational information system for the 21st century: The development of O*NET*. Washington, DC: American Psychological Association.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, *64*, 916-924. https://doi.org/10.1177/0013164404264848

Potosky, D. (2008). A conceptual framework for the role of the administration medium in the personnel assessment process. *Academy of Management Review*, *33*, 629-648. https://doi.org/10.5465/AMR.2008.32465704

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. [Short version of the Big Five Inventory (BFI-K): Development and validation of an economic inventory for assessment of the five factors of personality]. *Diagnostica*, *5*1, 195-206. https://doi.org/10.1026/0012-1924.51.4.195

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

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

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

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: https://www.R-project.org/.

Schäpers, P., Mussel, P., Lievens, F., König, C. J., Freudenstein, J.-P., & Krumm, S. (in press). *The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions*. Journal of Applied Psychology.

Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the Geneva emotion recognition test: An example of Rasch-based test development. *Psychological Assessment, 26,* 666-672. https://doi.org/10.1037/a0035246

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

Stevens, M. J., & Campion, M. A. (1996*). Teamwork–KSA information guide.* Arlington, VA: Vangent.

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

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117,* 440-463. https://doi.org/10.1037/a0018963

University of Cambridge Local Examinations Syndicate (2016). *Test your English - Adult learners.* Retrieved from: http://www.cambridgeenglish.org/test-your-english/adult-learners/

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

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

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*, 806-838. https://doi.org/10.1177/0011000006288127

Table 1

*Itemwise Comparison of the Number of Correct Answers in SJTs with Text Situation Descriptions/Text Responses and Omitted Situation Descriptions/Text Responses*

| Item | Cohen's *d* | $\eta^2$ | *t* | *Df* | *P* |
|---|---|---|---|---|---|
| 1 | 0.63 | 0.09 | 3.05 | 93 | < .003 |
| 2 | 0.47 | 0.05 | 2.31 | 93 | .012 |
| 3 | – 0.04 | < 0.01 | – 0.18 | 93 | .431 |
| 4 | 0.08 | < 0.01 | 0.39 | 93 | .348 |
| 5 | – 0.08 | < 0.01 | –0.38 | 93 | .353 |
| 6 | – 0.34 | 0.03 | –1.63 | 93 | .053 |
| 7 | 0.26 | 0.02 | 1.23 | 93 | .107 |
| 8 | 0.46 | 0.05 | 2.19 | 68.4 | .016 |
| 9 | 0.36 | 0.03 | 1.74 | 93 | .043 |
| 10 | 0.42 | 0.04 | 2.05 | 93 | .022 |
| 11 | 0.88 | 0.16 | 4.18 | 72.4 | < .003 |
| 12 | 0.04 | < 0.01 | 0.17 | 93 | .432 |
| 13 | 0.35 | 0.03 | 1.68 | 93 | .049 |
| 14 | 0.38 | 0.03 | 1.86 | 93 | .034 |
| 15 | 0.19 | 0.01 | 0.91 | 93 | .183 |
| 16 | 0.03 | < 0.01 | 0.12 | 93 | .451 |
| 17 | 0.47 | 0.05 | 2.28 | 93 | .013 |

*Notes.* One-sided *t* tests. Higher effect sizes reflect more correct answers to items with situation descriptions compared to items without situation descriptions.
* $p < .003$ (*p*-level adjusted to account for alpha inflation: *p*/number of tests = .05/17 = .003).

Table 2

*Itemwise Comparison of the Number of Correct Answers in SJTs with Video Situation Descriptions/Video Responses and Omitted Situation Descriptions/Video Responses*

| Item | Cohen's *d* | $\eta^2$ | *t* | *Df* | *P* |
|---|---|---|---|---|---|
| 1 | 0.06 | < 0.01 | 0.30 | 87 | .384 |
| 2 | 0.00 | < 0.01 | 0.00 | 87 | .500 |
| 3 | 0.38 | 0.03 | 1.79 | 87 | .039 |
| 4 | – 0.18 | 0.01 | – 0.86 | 87 | .198 |
| 5 | 0.68 | 0.10 | 3.18 | 87 | < .003 |
| 6 | 0.43 | 0.04 | 2.02 | 87 | .024 |
| 7 | 0.25 | 0.02 | 1.19 | 87 | .120 |
| 8 | 0.57 | 0.08 | 2.68 | 87 | .005 |
| 9 | 0.46 | 0.05 | 2.12 | 73.4 | .019 |
| 10 | 0.58 | 0.08 | 2.70 | 87 | .004 |
| 11 | 0.58 | 0.08 | 2.74 | 87 | .004 |
| 12 | 0.21 | 0.01 | 0.98 | 87 | .166 |
| 13 | 0.25 | 0.02 | 1.18 | 87 | .120 |
| 14 | 0.11 | < 0.01 | 0.51 | 87 | .306 |
| 15 | 0.05 | < 0.01 | 0.22 | 87 | .413 |
| 16 | – 0.05 | < 0.01 | – 0.25 | 87 | .401 |
| 17 | 0.70 | 0.11 | 3.25 | 75.6 | < .003 |

*Notes.* One-sided *t* tests. Higher effect sizes reflect more correct answers to items with situation descriptions compared to items without situation descriptions.
* $p < .003$ (*p*-level adjusted to account for alpha inflation: *p*/number of tests = .05/17 = .003).

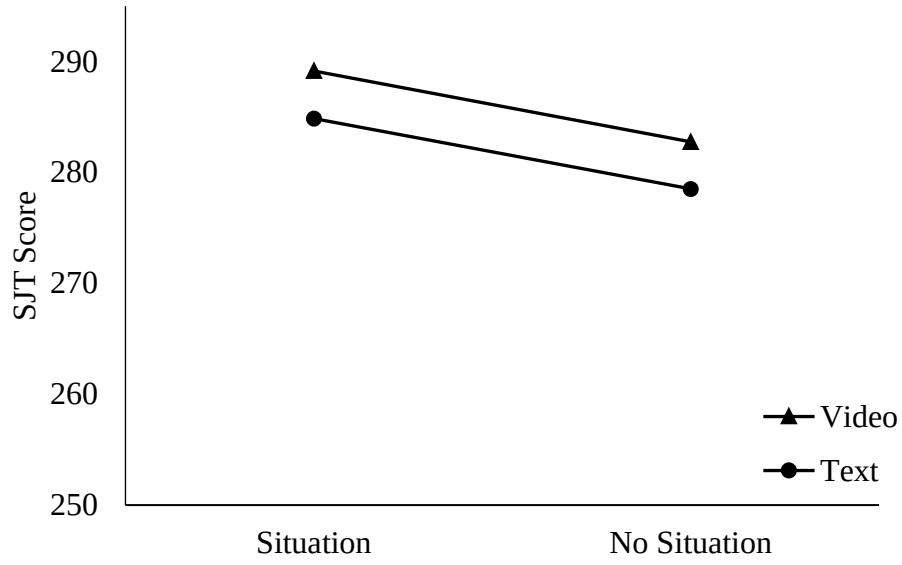


*Figure 1.* Comparison of SJT scores (with vs. without situation descriptions and video- vs. text-based modality).

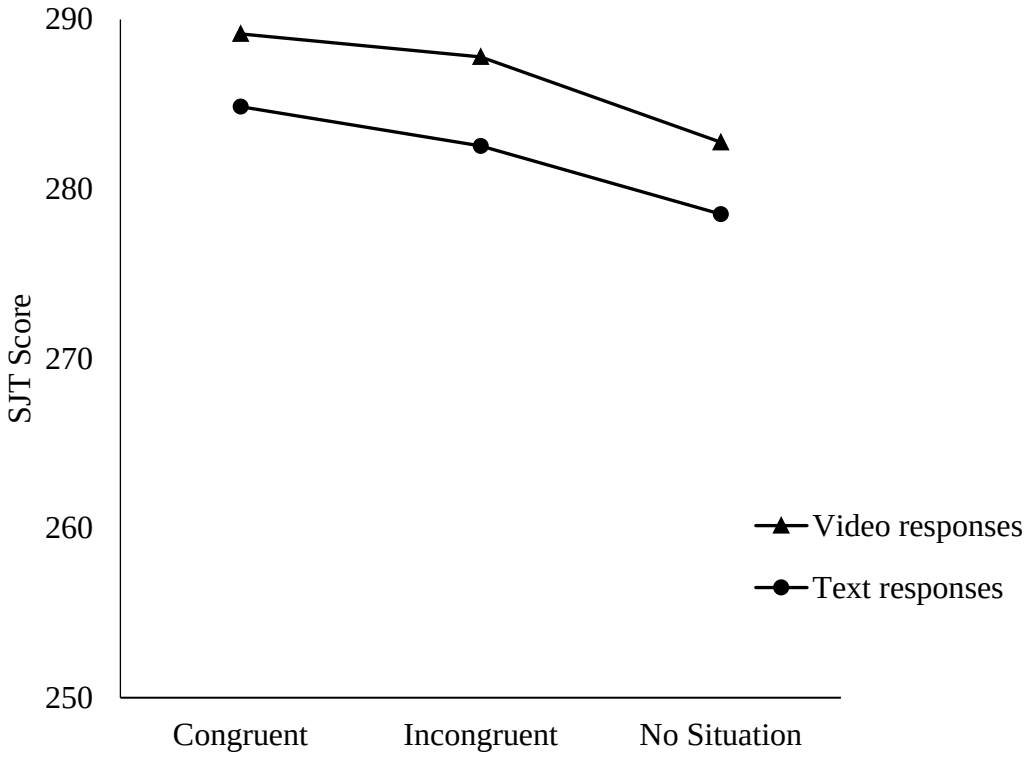


*Figure 2.* Effects of the factors *congruency* (congruent = same response and situation modality vs. incongruent = different response and situation modality vs. no situation = situations descriptions omitted) and *response modality* (video vs. text responses) on SJT scores.

## Appendix A

## Sample Multimedia SJT Item

| Sample video-based SJT item as administered in the SJT version with situation descriptions (Oostrom et al., 2012): | Sample video-based SJT item as administered in the SJT version without situation descriptions (Oostrom et al., 2012): |
|---|---|
| Situation Description:<br><br>*Two coworkers are supposed to work together on a project. However, the collaboration between the two coworkers is not going that well. One of the coworkers is complaining to the supervisor.*<br><br><u>Coworker</u><br>"Can I speak to you for a moment?"<br><br><u>Supervisor</u><br>"Of course"<br><br><u>Coworker</u><br>"I can no longer work this way! Peter is impossible to work with! He doesn't consult me on the project, when we have an agreement he doesn't stick to it, and he only does what he thinks is best. It doesn't work that way. I've tried talking to him about this problem, but he does not want to listen to me. I'm sorry, but I refuse to work with him on this project any longer!" | Please click on the reaction buttons (1 to 4) to score each reaction. |
| Possible Reactions:<br><br>a) <u>Manager</u><br>"Well, I can't just delegate this project to someone else. You can at least try working with him in a professional way. There are many colleagues who don't like each other, but are still capable of working together."<br><br>b) <u>Manager</u><br>"Well, that's impossible! We are all professionals and you cannot quit before finishing the project. I expect you will resolve this problem together. The project needs to be finished, you should understand that. | Possible Reactions:<br><br>a) <u>Manager</u><br>"Well, I can't just delegate this project to someone else. You can at least try working with him in a professional way. There are many colleagues who don't like each other, but are still capable of working together."<br><br>b) <u>Manager</u><br>"Well, that's impossible! We are all professionals and you cannot quit before finishing the project. I expect you will resolve this problem together. The project needs to be finished, you should understand that. |

c) Manager
"Oh my… what a hustle. I understand the problem has escalated and you can no longer work this way. To be honest, the project has to be finished on time. Maybe we can look for a solution. Do you have any idea how this problem can be resolved?"

d) Manager
"Too bad, that the collaboration is not going well… I propose that you tell me everything that's bothering you, so we can look for a possible solution for this problem. Is that alright with you?"

c) Manager
"Oh my… what a hustle. I understand the problem has escalated and you can no longer work this way. To be honest, the project has to be finished on time. Maybe we can look for a solution. Do you have any idea how this problem can be resolved?"

d) Manager
"Too bad, that the collaboration is not going well… I propose that you tell me everything that's bothering you, so we can look for a possible solution for this problem. Is that alright with you?"

## Appendix B

### Correlations of Video- and Text-based SJTs (With and Without Situation Descriptions) with Personality and Emotional Intelligence

| Measure | Video-based SJT: bivariate correlation with | | | Text-based SJT: bivariate correlation with | | |
|---|---|---|---|---|---|---|
| | SJT with situation descriptions | SJT without situation descriptions | Difference between correlations (*z*-score) | SJT with situation descriptions | SJT without situation descriptions | Difference between correlations (*z*-score) |
| *Personality* | | | | | | |
| Extraversion | .12 | .31* | -0.90 | -.16 | .09 | -1.20 |
| Agreeableness | .16 | -.02 | 0.81 | .05 | .14 | -0.40 |
| Conscientiousness | -.01 | .21 | -1.02 | .01 | .03 | -0.09 |
| Neuroticism | -.24 | .09 | -1.53 | .20 | .10 | 0.50 |
| Openness | .28 | .16 | 0.54 | .17 | .15 | 0.13 |
| *Emotional intelligence* | | | | | | |
| Emotional recognition | .04 | -.01 | 0.22 | .31* | .38* | -0.37 |
| Emotional understanding | .25 | .03 | 1.02 | .42** | .48** | -0.30 |
| Emotional management | .17 | .27 | -0.46 | .15 | .40** | -1.20 |

*Note.* $n_{\textit{video-based SJT, with situation descriptions}} = 48$, $n_{\textit{video-based SJT, without situation descriptions}} = 41$, $n_{\textit{text-based SJT, with situation descriptions}} = 50$, $n_{\textit{video-based SJT, without situation descriptions}} = 45$. Cronbach's Alpha for big five personality traits ranged from .51 to .70 and for the emotional intelligence tests from .28 to .64. * $p < .05$. ** $p < .01$.