Journal of Applied Psychology

0021-9010/14/$12.00 http://dx.doi.org/10.1037/a0037674

How "Situational" Is Judgment in Situational Judgment Tests?

Stefan Krumm
Freie Universität Berlin

Filip Lievens
Ghent University

Joachim Hüffmeier
Federal Institute for Occupational Safety and Health, Dortmund, Germany

Anastasiya A. Lipnevich
City University of New York

Hanna Bendels and Guido Hertel
University of Münster

Whereas situational judgment tests (SJTs) have traditionally been conceptualized as low-fidelity simulations with an emphasis on contextualized situation descriptions and context-dependent knowledge, a recent perspective views SJTs as measures of more general domain (context-independent) knowledge. In the current research, we contrasted these 2 perspectives in 3 studies by removing the situation descriptions (i.e., item stems) from SJTs. Across studies, the traditional contextualized SJT perspective was not supported for between 43% and 71% of the items because it did not make a significant difference whether the situation description was included or not for these items. These results were replicated across construct domains, samples, and response instructions. However, there was initial evidence that judgment in SJTs was more situational when (a) items measured job knowledge and skills and (b) response options denoted context-specific rules of action. Verbal protocol analyses confirmed that high scorers on SJTs without situation descriptions relied upon general rules about the effectiveness of the responses. Implications for SJT theory, research, and design are discussed.

Keywords: situational judgment test, knowledge, simulation, contextualization, validity

During a regular day at work, individuals make countless judgments on how to respond to the situations they encounter. This idea has been adopted to create a selection procedure called *situational judgment tests* (SJTs). The most frequently employed SJT format consists of written job-related situations that are presented with multiple-choice response options (McDaniel, Hartman, Whetzel, & Grubb, 2007; Motowidlo, Dunnette, & Carter, 1990). As evidenced by a recent meta-analysis (Christian, Edwards, & Bradley, 2010), SJT situations most often pertain to the construct domain of applied social skills.

Stefan Krumm, Institute of Psychology, Freie Universität Berlin; Filip Lievens, Department of Personnel Management, Work, and Organizational Psychology, Ghent University; Joachim Hüffmeier, Federal Institute for Occupational Safety and Health, Dortmund, Germany; Anastasiya A. Lipnevich, Department of Elementary and Early Childhood Education, Queens College and the Graduate Center, City University of New York; Hanna Bendels and Guido Hertel, Department of Psychology, University of Münster.

We would like to acknowledge the help of Thomas Rockstuhl, Mike Christian, and Britt De Soete, who made very valuable comments on earlier drafts of this submission. We also thank Christin Mattuschka, Roxana Junczyk, and Ayse Semiz for their help in collecting the data.

Correspondence concerning this article should be addressed to Stefan Krumm, Institute of Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany. E-mail: stefan.krumm@fu-berlin.de

When SJTs were reintroduced to the scientific community in 1990, there was an implicit assumption that they captured *context-dependent knowledge*. In fact, the label *situational judgment* implies that candidate responses are more effective when the specifics of each particular situation are taken into account and responses are adjusted to meet the situational demands. The other frequently used label to refer to SJTs, *low-fidelity simulations*, also attests to this contextualized knowledge perspective. Although low fidelity implies that SJT scenarios are not "veridical representations of the task stimulus" (Motowidlo et al., 1990, p. 640), the logic behind simulations stresses the importance of including job-related and realistic situations in SJTs to enable a contextualized judgment.

In recent years, another perspective has emerged, namely, that SJTs might also capture *context-independent knowledge* (Motowidlo, Crook, Kell, & Naemi, 2009; Motowidlo, Hooper, & Jackson, 2006). According to this perspective, SJTs tap mainly into general domain knowledge (general rules about the effectiveness of responses in a given domain). For instance, under this perspective, SJTs have been labeled as measures of *implicit trait policies*. As this label suggests, the knowledge needed to solve SJTs is deemed to be relatively context independent and applicable across a wide range of situations.

We view these two perspectives as two ends of a continuum, with most SJTs and their items falling somewhere in between. So far, little is known about the extent to which judgment in SJTs is contextualized or decontextualized. Similarly, we do not know which specific factors (e.g., type of constructs measured, type of

items, or type of sample) might moderate the extent to which judgment in SJTs is contextualized or decontextualized. Apart from implications for SJT theory, these questions are also important for SJT development (e.g., developing contextualized SJTs requires the input from subject matter experts). In the current article, we report results of a series of studies designed to shed light on these pressing questions by juxtaposing these two perspectives (SJT performance driven by context-dependent vs. context-independent knowledge). In the remainder of the introduction, we discuss the two SJT perspectives in more detail, thereby highlighting their importance for SJT theory and design.

Study Background

Situational Judgment Tests as Measures of Context-Dependent Knowledge

Simulations represent contextualized selection procedures that psychologically or physically mimic key aspects of the job (Lievens & De Soete, 2012). In accordance with this definition, SJTs confront applicants with a set of situations similar to those they might encounter later on the job and elicit their responses to these contextualized stimuli. Like other simulations (such as assessment center exercises or work samples), SJTs build on the notions of point-to-point correspondence with the criterion (future job situations and behavior) and behavioral consistency (Lievens & De Soete, 2012; Thornton & Cleveland, 1990). However, in light of cost concerns, the vast majority of SJTs adopt a low-fidelity format in simulating the situations and responses. That is, SJTs typically present job-related situations in the form of written descriptions and require respondents to react to them by picking a response to the situation from a list of responses (McDaniel et al., 2007; Weekley, Ployhart, & Holtz, 2006).

As this perspective emphasizes contextualization, the situation descriptions are considered to play a key role in SJTs. These descriptions aim to simulate job contexts, thereby influencing candidates' situation perception, subsequent response selection, and response effectiveness. Thus, the situation descriptions in SJTs are intended to provide contextualization to candidates so that they can imagine themselves in the situation and make well-thought-out judgments among alternative ways of responding (Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000). One implication of this rationale is that without contextual information, candidates cannot solve the item. This also illustrates that this SJT perspective is grounded in interactionism. That is, it assumes that candidates' behavioral response selection is contingent upon how they perceive and construe the stimuli (job-related situations). The latter view aligns well with recent interactionist theories that posit reciprocal influences between situation perception and behavioral response selection (Campion & Ployhart, 2013; Mischel & Shoda, 1995).

This traditional SJT perspective has important implications for SJT design as it suggests that it makes sense to increase the level of contextualization and realism in SJT situations. In fact, over the years, various technologies have been proposed as alternatives to the traditional written SJT presentation format. A well-known example consists of using video-based, instead of written scenarios, as SJT stimuli (e.g., Lievens & Sackett, 2006). Other examples are the use of item branching or three-dimensional animation for presenting the situations in a way that bears closer resemblance to actual job situations (e.g., Kanning, Grewe, Hollenberg, & Hadouche, 2006; Tippins & Adler, 2011).

Situational Judgment Tests as Measures of Context-Independent Knowledge

In recent years, another perspective has emerged that describes SJTs as measures of general knowledge that is seen as more context independent. In a series of articles, Motowidlo and colleagues (Motowidlo & Beier, 2010; Motowidlo et al., 2006) have provided the conceptual foundation for this perspective. According to these researchers, general domain knowledge can be defined as general rules about the utility of behavioral acts across a wide range of situations in a specific domain. The more general this knowledge is, the more context independent and the more broadly applicable it is across a wide range of situations. That is why this general domain knowledge is also referred to as *implicit trait policies* (Motowidlo et al., 2006), which are inherent and personality-dependent beliefs about the general effectiveness of a trait expressed. The origins of general domain knowledge do not stem from specific job experiences. Instead, general domain knowledge originates from fundamental socialization processes (parents, schooling, and so forth) and personal dispositions.

Other scholars have suggested concepts that are somewhat similar to implicit trait policies such as practical intelligence (cf. Wagner & Sternberg, 1985) or heuristic decision making (Gigerenzer & Gaissmaier, 2011). Practical intelligence, however, is distinct from implicit trait policies as it refers to individuals' tendencies in dealing with context-specific problems. For example, Wagner and Sternberg adopted Neisser's (1976) definition to refer to practical intelligence as "responding appropriately in terms of one's long-range and short-range goals, *given the actual facts of the situation as one discovers them*" [italics added] (p. 437). Heuristic decision making is a broader concept than implicit trait policies as it refers to strategies that ignore "part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods" (Gigerenzer & Gaissmaier, 2011, p. 454). Hence, implicit trait policies represent strategies as part of the broader concept of heuristic decision making—strategies that have guided recent theorizing in the realm of SJTs (Motowidlo et al., 1990) and were adopted in the current article.

The conceptualization of SJTs as measures of relatively context-independent knowledge has also key implications for SJT design. If SJTs measure general domain knowledge, it seems less vital to invest in elaborate contextualized situation descriptions as it is often recommended by the aforementioned traditional approach. Instead, this perspective conceptually guides efforts to streamline SJTs. One example is the use of more generic situation descriptions (e.g., "A customer is rude to you; what's the best response?"). Such more generic SJT items require knowledge that is typically applicable across a broad range of situations. Another example is the use of single-response SJTs (Crook et al., 2011; see also Motowidlo et al., 2009; Motowidlo, Martin, & Crook, 2013). In single-response SJTs, item stems are deleted to reduce development costs. Instead, each item consists of a couple of sentences describing one critical incident (i.e., a response in a particular

situation). Candidates are asked to rate the effectiveness of the response. The emphasis is placed on assessing candidates' judgment of the effectiveness of the behavior shown in the response option. Crook et al. (2011) found single-response SJTs to be valid predictors of performance.

Hypotheses

Taken together, two perspectives about the determinants of SJT performance seem to have emerged in the literature. In this article, we contrasted these two perspectives by creating two SJT conditions: In one condition, participants received the SJT with both situation descriptions and response alternatives. In the other condition, the situation descriptions were removed, and the individuals received only the response alternatives.

Comparing scores across those two conditions permits examination of whether successful performance on SJT items requires the availability of the contextual information provided by the item stem. If performance on an SJT item is significantly impaired when the contextual information is absent, this indicates that the SJT item taps rather into context-dependent knowledge as this shows that test takers need to have an accurate portrayal of the situation to make a thoughtful judgment about possible responses. Framed in terms of item generation theory (Irvine & Kyllonen, 2002; Lievens & Sackett, 2007), this would show that the situation description is a radical (i.e., a key characteristic that determines performance) instead of an incidental (i.e., a superficial item unrelated to performance) feature. Conversely, if performance on the SJT item is not significantly affected, this would suggest that the SJT item captures rather context-independent knowledge (general domain knowledge). In that case, the knowledge needed to solve the item is relatively independent of the situation because the effective response is applicable to many situations. In that case, the situation description serves as an incidental instead of a radical feature.

In short, according to the context-dependent knowledge perspective, there will be a significant difference in scores on SJT items with and without situation descriptions (Hypothesis 1a) as this perspective stresses that people need to have an accurate portrayal of the situation to make a thoughtful judgment about possible responses. Conversely, the context-independent knowledge perspective posits that the difference between the two conditions will be nonsignificant (Hypothesis 1b) as this perspective emphasizes that SJT items tap into knowledge that might be relatively independent of the situations. This leads to the following competing hypotheses:

> *Hypothesis 1a:* There will be a significant difference in scores on SJT items with and without situation descriptions. Scores on SJT items without situation descriptions will be significantly lower than scores on SJT items with situation descriptions.

> *Hypothesis 1b:* There will not be a significant difference in scores on SJT items with and without situation descriptions.

Overview of Objectives and Studies

The objectives of this research were twofold. First, this research was the first to contrast these two perspectives about the knowledge determinants of SJT performance by comparing SJT performance across the two aforementioned conditions. This investigation might shed light on how "situational" judgment really is in SJTs. Second, we sought a better understanding as to when one of the two knowledge perspectives underlies SJT performance. Thus, the second research objective was on illuminating factors that might moderate whether SJT item performance is contextualized or decontextualized.

Three studies are presented. In Study 1, we started by implementing our experimental design with the two conditions using one of the most popular commercially available SJTs (i.e., the Team Knowledge–KSA Test; Stevens & Campion, 1996; see also O'Neill, Goffin, & Gellatly, 2012). In Study 2, we provided a more stringent comparison of the two perspectives by varying the construct domain captured by the SJT, thereby also testing the generalizability of the findings of Study 1 across other constructs targeted by SJTs. Finally, in Study 3, we explored underlying thought processes of participants completing an SJT without situation description by gathering their verbal protocols.

Study 1

Method

Participants and procedure. Study 1 included 436 individuals (35.8% male). To increase the generalizability of the findings, we used individuals from two samples. The first sample comprised 257 students (28% male) from various academic fields. On average, students were 23 years old ($M = 22.56$, $SD = 3.16$, range = 18–42). Almost all students (99.6%) had some professional experience, which they had gained in student jobs.

The second sample consisted of 179 working people (47% male). On average, participants were 44 years old ($M = 44.11$, $SD = 11.38$, range = 21–70) and worked in various occupational fields. The sample included individuals who held jobs with and without leadership responsibilities (19% and 81%, respectively). The vast majority of working people (88%) reported that their jobs required teamwork. Both working people and students were recruited via postings (on websites or in newsletters) or were actively approached (mostly via e-mail) by several research assistants. They participated voluntarily and anonymously. All participants were debriefed about the aim of the study after completing the SJT.

The assessment was conducted in proctored and unproctored sessions. In the proctored test sessions (about 40–50 min long), groups of between four and 15 individuals were tested at the same time. In the unproctored test setting, participants were given an envelope containing the instructions, the SJT test version, and one page of demographic questions. Participants were asked to complete the SJT at home and to return it in the envelope to the test administrator. Participants were randomly assigned to these test conditions. Due to organizational constraints, proctored test sessions were not possible with working people. Note that we included the type of test administration (proctored vs. unproctored) in our analyses as it could impact SJT results (cf. Potosky & Bobko, 2004).

Experimental design. As noted earlier, our basic experimental design was a between-subjects design consisting of two conditions (SJT version: with vs. without situation descriptions). The

Teamwork–KSA Test (Stevens & Campion, 1994, 1996, 1999) served as SJT. This test is an established and widely used commercially available SJT (for a review, see O'Neill et al., 2012).[1] In line with typical SJTs, the 35 items of the Teamwork–KSA Test consist of situation descriptions depicting realistic work situations followed by four response options. Situations described in the Teamwork–KSA Test refer to conflict resolution, collaborative problem solving, communication, goal setting and performance management, and planning and task coordination. In the original version, a knowledge-based response instruction was adopted, instructing the test takers to select the response option that reflects the most effective reaction in a given situation. Participants received either 1 or 0 points per each item, depending on whether they chose the most effective response option or not. Test scores were computed by summing up participants' points across all items. An expert scoring procedure conducted by the authors of the Teamwork–KSA Test determined the most effective response alternative per each item. Consistent with our experimental design, we created a stemless version of this SJT by omitting the situation descriptions of the original Teamwork–KSA Test. Reliability estimates as obtained in the present sample are reported in the Preliminary Analyses section. A sample item as it was formulated in both experimental conditions can be found in Appendix A.

Apart from these two basic forms of the Teamwork–KSA Test that reflected the basic design of the study and were related to our main research objective, we also varied the response instruction used to verify whether the traditional knowledge-based format of the Teamwork–KSA Test did not favor one of the perspectives. In fact, prior meta-analytical research has shown that the use of knowledge-based instructions makes the SJT more similar to an actual knowledge test (what is the "right thing" to do?) and a measure of maximal performance (McDaniel et al., 2007), whereas the use of behavioral tendency response instruction in SJTs (i.e., "What would you do in this situation?") measures rather typical performance. So, in both aforementioned versions of the Teamwork–KSA Test (with and without situation descriptions), we also modified the traditional knowledge-based response instruction to a behavioral tendency response instruction. In total, this led to four different forms of the Teamwork–KSA Test, to which participants were randomly assigned (in the condition without situation descriptions, $n = 110$ received a knowledge-based and $n = 100$ a behavioral tendency instruction; in the condition with situation descriptions, those numbers were $n = 105$ and $n = 121$, respectively).

Results

Preliminary analyses. We conducted preliminary analyses to rule out several alternative explanations for possible differences across the SJT versions with and without situation descriptions. First, we examined whether the scores on the two SJT versions differed in reliability. The internal consistency reliabilities were generally low (split-half reliabilities around .50), which is in line with the SJT literature (see O'Neill et al., 2012, for similarly low internal consistency reliabilities for this particular SJT test; see Catano, Brochu, & Lamerson, 2012, as well as Kasten & Freund, 2013, for meta-analyses of internal consistencies in SJTs). It should be noted that testing for differences across these reliability estimates (see Cohen, Cohen, West, & Aiken, 2003) did not yield significant results across conditions (z scores from 0.03 to 0.48, all p values $> .05$).

Second, we performed multiple group measurement invariance analyses using SPSS Amos Version 21 to examine whether the measurement structure underlying the Teamwork–KSA Test scores remained stable in both conditions (with and without situation descriptions). Prior to testing for measurement invariance, we specified various models. To this end, we specified the same models as in O'Neill et al.'s (2012) comprehensive psychometric evaluation of this specific SJT. That is, the following models were specified and tested: a single-factor model, a two-factor model (interpersonal, self-management), and a five-factor model (conflict resolution, communication, collaborative problem solving, goal setting and performance management, and planning and task coordination).

Consistent with findings reported by O'Neill et al. (2012), these measurement models showed generally a poor model fit (see Appendix B). The best fit was obtained for the two-factor model, χ^2 (559) = 694.29, $p < .01$, root-mean-square error of approximation (RMSEA) = .034, 90% confidence interval (CI) for RMSEA = .025, .042, standardized-root-mean square residual (SRMR) = .068, which still showed a low comparative fit index (CFI = .54). The low CFI value is in line with the finding that the internal consistency reliability of the Teamwork–KSA Test scores is low (cf. Kline, 2004). All other fit indices fell into an acceptable range as recommended by several researchers (Beauducel & Wittmann, 2005; Byrne, 1989; Hu & Bentler, 1999): $\chi^2/df < 2$, RMSEA $\leq .08$, and SRMR $\leq .09$. We used the two-factor model as baseline model for our assessment of more restrictive forms of measurement invariance (invariance of factor loadings and invariance of variances and covariances, see Byrne, 2004). These subsequent analyses showed that there was evidence of factor invariance and variance/covariance invariance across the two SJT versions (as indicated by insignificant increases in chi-square values).

In sum, although measurement invariance was found across the two SJT versions, some fit indices were low, indicating that the specified structure of our model cannot be considered a better fit than the independence model, which assumes no correlation between items (cf. Kline, 2004). In light of these findings and the fact that the hypotheses were formulated at the SJT item level, we conducted all analyses at the SJT item level. This approach also takes into account that some items might rely more on context-(in)dependent knowledge than others.

Hypothesis tests. According to Hypothesis 1a, scores on SJT items without situation descriptions would be significantly lower than scores on SJT items with situation descriptions, whereas Hypothesis 1b did not posit such significant differences on the items. Basically, this first hypothesis deals with whether the difficulty of the items differed across conditions. As an overall test, we conducted a one-factorial analysis of variance (ANOVA) to examine whether the independent variable SJT version (with or without situation descriptions) captured a significant part of the

[1] The current study is not intended to evaluate this SJT in particular. The test was chosen based on its prominence in the academic literature, its diffusion in the field, and its features that make it a good representative of SJTs in general.

variation in item difficulty across the items. So, in this ANOVA, the item was the unit of analysis, and the proportion of correct solutions served as dependent variable. This proportion was calculated separately for each condition the item was administered in and reflected the number of participants per condition who solved an item correctly divided by the total number of participants per condition. We used the proportion of correct solutions to account for the slightly unequal number of participants in the experimental conditions. The ANOVA did not yield a significant effect, $F(1, 68) = 3.275$, $p = .075$, $\eta = .214$, lending support to Hypothesis 1b. Acknowledging the moderate effect size obtained, we decided to follow up on this result with *t* tests per each item. In addition to reporting uncorrected values, we also accounted for alpha-error inflation by adjusting the critical alpha level with a Bonferroni correction (Cabin & Mitchell, 2000). Results of the respective *t* tests are shown in Table 1. When uncorrected *p* values are used, 19 out of 35 items (54%) were identified as resulting in significantly more correct answers (i.e., these items were easier to solve) when situation descriptions were given. When corrected *p* values were used, 10 out of 35 items (28.6%) yielded significantly more correct answers when situation descriptions were given. The average effect size across the items was 0.18 (range = 0.25–0.80). In sum, using the conservative alpha value as yardstick, only 10 items provided support for Hypothesis 1a (the context-dependent knowledge perspective). Conversely, 25 items provided support for Hypothesis 1b (the general domain knowledge perspective), which is in line with the result of the overall ANOVA reported earlier.

Table 1
Itemwise Comparison of the Number of Correct Answers Depending on the Availability of Situation Descriptions (Study 1)

Item	Word count of situation description	Glass δ	*t*	*df*	*p*
1	17	0.08	0.79	434	.215
2	14	0.03	0.32	433	.368
3	32	0.33	3.10	410.3	*
4	17	0.41	4.15	424.3	*
5	18	0.32	3.32	428.4	*
6	13	0.22	2.04	397.5	.021
7	16	−0.11	−1.16	431.0	.123
8	37	0.06	0.60	433	.276
9	32	−0.10	−1.06	433	.145
10	23	0.68	5.68	366.6	*
11	22	0.08	0.81	433	.210
12	55	0.78	10.47	331.0	*
13	17	0.10	1.03	426	.152
14	34	0.50	5.94	405.9	*
15	38	−0.06	−0.60	432	.276
16	6	0.13	1.39	428.9	.083
17	28	0.47	4.87	427	*
18	16	0.23	2.47	430.6	.007
19	29	0.05	0.49	434	.313
20	31	0.27	2.69	421.3	.004
21	29	0.09	0.92	431	.180
22	45	−0.01	−0.07	433	.474
23	44	0.17	1.73	429.8	.043
24	21	0.75	8.87	415.4	*
25	32	−0.25	−2.41	409.9	.009
26	24	−0.12	−1.23	432.3	.111
27	18	−0.15	−1.51	434	.065
28	38	0.45	5.45	402.7	*
29	9	0.80	10.76	317.5	*
30	19	0.26	2.69	432	.004
31	15	0.24	2.55	433.9	.006
32	18	−0.21	−2.12	427.4	.018
33	22	0.12	1.14	419.5	.128
34	17	−0.17	−1.84	433	.034
35	40	0.01	0.06	430	.475

Note. One-sided *t* tests. Degrees of freedom vary due to the use of *t* tests for homogeneous and heterogeneous variances. Higher effect sizes reflect more correct answers on items with situation descriptions compared with items without situation descriptions.
* $p < .00143$ (*p* level adjusted to account for alpha inflation: *p*/number of tests = .05/35 = .00143).

As for the practical effect of the significant item level differences, we found a mean difference of 2.97 (on 35 points) between the two SJT versions on the overall score, indicating that on average about three items more were solved in the SJT version employing situation descriptions. The average score of test takers who completed the SJT items without situation descriptions was 17.67 ($SD = 3.36$), showing that the stemless SJT items still made sense to them and that they were able to complete them. The mean score on the SJT items with situation descriptions was 20.64 ($SD = 4.04$).

Additional analyses. To identify moderators of differences between items with and without situation descriptions, we conducted various additional analyses. First, we extended the ANOVA reported previously and conducted a 2 (SJT version: with vs. without situation descriptions) × 2 (sample: students vs. working people) × 2 (instruction: knowledge-based vs. behavioral tendency response instruction) × 2 (test administration: proctored vs. unproctored) ANOVA. No main effects associated with these additional factors were found: sample, $F(1, 408) = 1.70$, $p = .193$, partial $\eta^2 = .004$; response instruction, $F(1, 408) = 1.81$, $p = .179$, partial $\eta^2 = .004$, administration; $F(1, 408) = 0.01$, $p = .973$, partial $\eta^2 = .000$. Additionally, no significant interactions of the SJT version factor with any of these extra factors was observed, indicating that these factors did not act as moderators of differences between SJTs with and without situation descriptions. All other higher order interactions were also not significant.

Moreover, we scrutinized the 10 items that produced significant differences with the Bonferroni correction. In particular, we considered various explanations for why these differences might have occurred. One possible explanation is related to the length of the item stem and item responses. For instance, it is plausible that the item length is positively correlated with the amount of contextualization so that the context matters only in longer items. However, length of items or response options were not related to differences between the two SJT versions (e.g., correlations between effect sizes given in Table 1, i.e., Glass δ, and word count of situation description: $r = .07$).

As a second potential explanation, we examined whether more difficult items might provide significant differences between the two conditions as the solution of such items require more fine-grained contextual information. To this end, we correlated the item difficulty properties (as obtained from SJT items with situation descriptions) with the effect sizes given in Table 1 (Glass δ). No systematic effect was found ($r = .05$).

Finally, we inspected the content of the response options of the 10 items that differed significantly when administered with or without situation descriptions. Two experts in the domain of teamwork who were blind to the results of the current study indepen-

dently coded whether the responses of these items mentioned general versus more context-specific rules of action (interrater agreement of $\kappa = .62$; discrepant codings were then discussed and resolved by the coders). It was observed that those 10 items, which produced significant differences, tapped into more context-specific rules of action (e.g., deciding how to involve the team when urgently preparing training material). Indeed, items with context-specific rules of action produced significantly larger effects in our main analysis, $t(33) = 3.18$, $p < .01$, Glass $\delta = 1.51$ (using the standard deviation of items with unspecific responses; Glass, McGraw, & Smith, 1981).

Discussion

The first key result of Study 1 was that it did not make a significant difference whether the situation description was included for between 46% (when no correction was applied to the alpha level for making multiple comparisons) and 71% of the items (when the alpha level was corrected) of a team knowledge SJT. This means that the context-independent perspective applies for between about 50% and 70% of the items of this team knowledge SJT. Given the substantial manipulation executed (i.e., the SJT items were virtually "decapitated" by removing their item stems), one might have anticipated much lower percentages. In addition, contrary to recent SJTs of Motowidlo and colleagues (Motowidlo & Beier, 2010; Motowidlo et al., 2006), the team knowledge SJT was not specifically developed to assess context-independent implicit trait policies (general domain knowledge). This is another reason why lower percentages might have been expected. As noted earlier, an assumption underlying this particular SJT—and most SJTs in general—is that they are contextualized methods wherein the simulated work-related situations play a crucial role for choosing the correct response. Hence, in light of the fact that "situational" is an integral part of the term "SJT," it is striking that so many items can be solved without the situation description.

Study 1 also started to shed some light on factors that might moderate this result. The type of response instruction did not emerge as a significant moderator as the findings generalized across knowledge and behavioral tendency instructions. Probably, the absence of the situation description makes the SJT instruction manipulation moot. In addition, the type of sample did not make a difference. Similar results were obtained across students and people who had actual teamwork experience in organizations. Given their actual teamwork experience in business organizations, one might have expected the working people to benefit more from the presence of situation descriptions (or to miss their absence more) as the context-dependent SJT perspective posits that such descriptions are needed to provide a contextual and realistic background in making judgments. One explanation for the correspondence between the student and working people results is that reliance on general interpersonal knowledge (instead of fine-grained contextualized teamwork knowledge) suffices to complete this team knowledge SJT. However, it is also possible that the working people were less motivated to participate in the study.

Regarding item characteristics as moderators, there was no evidence that item length or difficulty moderated the results. However, there was some evidence that the type of response option moderated the results. In case no situation description was given, SJT performance in both the student and working people samples was lower for those 10 items that had response options denoting context-specific courses of action compared with general courses of action.

Given that this was the first study to examine how "situational" SJTs are, these results beg for replication. As results of Study 1 pertain only to a team knowledge SJT, we do not know whether these results are due to the content domain assessed (team knowledge) or to the SJT method. Indeed, it might be that the items of this specific SJT primarily capture context-independent knowledge. Therefore, to take the construct-method distinction into account and to make a more generalizable comparison of the two SJT perspectives, we conducted a second study wherein we varied the construct assessed with the same measurement method (SJT) and additionally controlled for participants' test motivation.

Study 2

The main objective of Study 1 consisted of examining to what extent judgment on SJT items is either contextualized or decontextualized. We also began investigating potential measurement-related moderators (e.g., response instructions and item characteristics). Study 2 takes this one step further by focusing on construct-related moderators. That is, we posited that the content domain captured by the SJT might be a key moderator of whether SJT performance is determined by context-independent knowledge versus context-dependent knowledge. SJTs are measurement methods that have the potential to assess a variety of constructs (Christian et al., 2010). As already mentioned, SJTs might therefore be placed on a continuum, with some SJTs measuring rather context-independent knowledge and others being situated on the context-dependent knowledge side. More generally, Christian et al. (2010) classified the various constructs assessed by SJTs in a taxonomy consisting of three broad categories: job knowledge and skills (e.g., SJTs assessing pilot judgment or knowledge relevant for firefighters), applied social skills (e.g., teamwork or leadership SJTs), and basic personality tendencies (e.g., SJTs gauging Big Five personality dimensions or facets such as integrity).

We expected that the first category of SJTs (job knowledge and skills) would be placed more along the context-dependent side of the continuum, whereas SJTs assessing applied social skills and basic personality tendencies would be more context-independent. This assumption is in line with the definitions of these three different construct domains. For instance, basic personality tendencies are typically defined as bundles of behaviors that generalize across situations (e.g., Fleeson & Gallagher, 2009). Likewise, social skills are broadly defined as acting wisely in larger classes of social situations (e.g., Thorndike & Stein, 1937; Topping, Bremner, & Holmes, 2000). Conversely, the term *job knowledge and skills* stresses this category's specificity for a given job. Huffcutt, Conway, Roth, and Stone (2001) referred to this category as capturing experience in a specific job. They subsumed, for instance, product and technical knowledge under this category. Hence, we anticipated that the absence of contextual information (situation descriptions) would impair performance on SJT items related to job-specific knowledge more strongly than performance on SJT items assessing applied interpersonal skills or basic personality tendencies.

To further illustrate this, aviation pilots, for instance, might find it difficult to answer an SJT item measuring how to conduct a landing procedure without situational information (visibility, weather, position, length of the runway, and so forth). The lack of contextual information might be less of a roadblock for SJT items assessing more general personal tendencies, such as applied social skills and personality. So, this example shows that the underlying measured construct (skill to land a plane vs. teamwork) might play an important role in the debate about SJT judgment being either contextualized or decontextualized. At a more molecular level, results of Study 1 are suggestive of this. Recall that Study 1 revealed the type of response options (denoting context-specific vs. more generic courses of action) to be associated with whether there was a difference on SJT performance with and without situation descriptions. As response options are essentially an operationalization of the construct(s) targeted by the SJT, it can be assumed that on a more general level, the nature of the context specificity of the construct itself serves as a moderator of the context dependency of SJT performance. In sum, this leads to the following hypothesis:

> *Hypothesis 2:* The difference in scores on SJT items with and without situation descriptions will be moderated by the content domain captured by the SJT, such that this difference will be more pronounced for SJT items assessing job knowledge and skills than for SJT items assessing applied social skills or basic personality tendencies.

Method

Participants. We decided to collect an appropriately experienced sample related to the job knowledge and skills SJT (which was about pilot judgment). Therefore, the sample consisted of pilots. In addition, it made also sense to administer the other SJTs of Study 2 (about teamwork and integrity) to pilots as these construct domains are also relevant in their job. So, our sample consisted of 559 pilots who were approached through postings and mailings by airlines, airline unions, and foundations for pilots. Seven pilots were excluded from further analyses as they reported that they had no flight experience. The remaining sample reported about 16 years of average flight experience ($M = 15.91$, $SD = 12.14$, range = 0–57 years) and 7,487 flight hours ($M = 7{,}487.82$, $SD = 22{,}171.74$, range = 8–506,000 flight hours). All of them either held one or more professional flight certificates or were enrolled in a professional flight training program. Study participation was voluntary and anonymous. However, as a prerequisite for the study to be posted by airlines and airline unions, we were not allowed to collect data on participants' age and sex. Participants were debriefed about the aim of the study after they had completed the SJTs.

Design and procedure. Each participant completed items of three different SJTs (SJT domain: job knowledge and skills, applied social skills, and personality). Similar to Study 1, participants were randomly assigned to one of two conditions: In one condition, they received the SJT items with both situation descriptions and response alternatives. In the other condition, the situation descriptions were removed, and they received only the response alternatives. Given that in Study 1, no interaction effects of the SJT version factor (with vs. without situation description) with the response instruction factor and the test environment factor, respectively, were found, all participants in Study 2 received a behavioral tendency response instruction and were tested in an unproctored environment. Prior to taking each one of the three SJTs, participants were asked to indicate their job experience in the respective domain. At the end of the assessment, participants completed a test motivation questionnaire.

Situational judgment tests (SJTs). Three different SJTs were administered (either with or without situation descriptions) to cover the three broad content domains of SJTs (cf. Christian et al., 2010): (a) job knowledge and skills, (b) applied social skills, and (c) basic personality tendencies. For each of the domains, a representative SJT was chosen for inclusion in the current study. To keep test duration and motivation at an acceptable level, we randomly selected 10 items from each full SJT version. Reliability estimates of SJTs are discussed in the Results section. SJTs were administered in English in the sequence of their appearance in the following.

Pilot Judgment Inventory (Hunter, 2003). The Pilot Judgment Inventory was developed to capture general aviation pilots' decision making in realistic flight scenarios. Therefore, Christian et al. (2010) coded this SJT into the domain of job knowledge and skills. The 39 items of this test confront participants with short descriptions of flight scenarios, which are followed by four response alternatives. Participants receive either 1 or 0 points per each item, depending on whether they select the most effective response option or not. Test scores are computed by summing up participants' points across all items. An expert scoring procedure defined the most effective response alternative per each item. In prior research, the Pilot Judgment Inventory showed convincing validity evidence (e.g., higher SJT scores were related to lower hazardous flight events; see Hunter, 2003). To verify that the 10 randomly sampled SJT items still captured job knowledge and skills, we also administered these items to 190 individuals with no prior experience as a pilot. Results showed that lay people ($M = 5.30$, $SD = 1.71$) scored significantly lower than the pilots in our sample ($M = 6.1$, $SD = 1.55$), $t(303.11) = 6.09$, $p < .001$, Glass $\delta = .55$. A sample item from Hunter (2003, p. 377) follows:

> You are flying an "angel flight" with a nurse and noncritical child patient to meet an ambulance at a downtown regional airport. You filed visual flight rule: it is 11:00 p.m. on a clear night when, at 60 nm out, you notice the ammeter indicating a battery discharge and correctly deduce the alternator has failed. Your best guess is that you have from 15 to 30 min of battery power remaining. You decide to:
>
> (a) Declare an emergency, turn off all electrical systems, except for 1 NAVCOM and transponder and continue to the regional airport as planned.
>
> (b) Declare an emergency and divert to the Planter's County Airport, which is clearly visible at 2 o'clock at 7 nm.
>
> (c) Declare an emergency, turn off all electrical systems, except for 1 NAVCOM, instrument panel lights, intercom, and transponder, and divert to the Southside Business Airport, which is 40 nm straight ahead.
>
> (d) Declare an emergency, turn off all electrical systems, except for 1 NAVCOM, instrument panel lights, intercom, and transponder, and divert to Draper Air Force Base, which is at 10 o'clock, at 32 nm.

Following the rationale reported in Study 1 and our experimental design, we built a stemless version of the Pilot Judgment Inventory by omitting the situation descriptions. In this form, participants were prompted to answer the statement "You decide to . . ." but did not receive the situation description. Interviews with subject matter experts confirmed the suitability of this SJT for European commercial and private pilots.

Teamwork–KSA Test (Stevens & Campion, 1994, 1996, 1999). Administration of the Teamwork–KSA Test was identical to the procedure described in Study 1, with the exception that 10 randomly chosen items were presented, and only the behavioral tendency instruction was used.

Employee Integrity SJT (Becker, 2005). The Employee Integrity SJT is used to assess attributes of high-integrity employees, who possess, for example, benevolence, honesty, and productivity. To this end, the test presents 20 prototypical SJT items describing work-related scenarios in which employees may display more or less integrity-related behavior by choosing one of four response alternatives. Participants receive either 1 or 0 points per item, depending on whether they select responses pertaining to integrity-related behavior. The coding of responses was done through empirical scoring (Becker, 2005). The Employee Integrity SJT showed substantial correlations with several job performance criteria, thus confirming its validity. A sample item from Becker (2005, p. 229) follows:

> You're a new clerk in a clothing store and are being trained by Angie, a veteran employee. She quietly tells you that because employees are paid minimum wage, most people sometimes take home clothes for themselves. Employees who don't are considered dumb and arrogant. At closing time, Angie hands you a scarf to take home. Which of the following would you most likely do?
>
> (a) Take home the scarf and keep your mouth shut.
>
> (b) Take home the scarf, but return it to the shelf later without letting other employees see you.
>
> (c) Politely tell Angie that you don't need any more scarves.
>
> (d) Tell Angie that you don't want to take home any clothes, now or ever.

In the condition in which participants received no situation descriptions, we simply asked, "What would you do?" (without giving any further information about the situation). In some cases, slight modifications of this question were necessary, for example, "Which of the following would you most likely say?"

Test motivation. Five items from the Test Attitude Survey (TAS; Arvey, Strickland, Drauden, & Martin, 1990) were used to assess test motivation. Those five items were chosen from the Motivation Scale of the TAS based on their psychometric properties (see McCarthy & Goffin, 2003). A sample item is "I wanted to do well on these tests." Participants responded to these items on a 5-point Likert scale ranging from *strongly disagree* to *strongly agree*. The internal consistency reliability of ratings on these five test motivation items was satisfactory (Cronbach's $\alpha = .73$).

Professional experience. Before taking items from the Pilot Judgment Inventory (Hunter, 2003), participants were asked to indicate their flight experience (e.g., number of flight hours, years of experience as a pilot). Prior to responding to Teamwork–KSA Test items (Stevens & Campion, 1996), participants' teamwork experience was assessed (e.g., years of experience in working in a team, overall rating of teamwork experience). As the Employee Integrity SJT (Becker, 2005) taps into the domain of basic personality tendencies (Christian et al., 2010), no specific work-related experience was assessed.

Results

Preliminary analyses. We first examined participants' test motivation. Generally, test motivation was above the center of the scale ($M = 3.64$, $SD = 0.70$). A t test for independent samples revealed that test motivation differed between the two groups which either received SJTs with or without situation descriptions, $t(544) = 2.18$, $p < .05$. Notably, pilots who did not receive the situation descriptions reported being *more* motivated. The size of this effect was small ($\delta = 0.19$). Controlling for test motivation in additional analyses did not change the pattern of results reported in the following.

Similar to Study 1, we also compared the internal consistency reliabilities of the SJT scores between the two conditions (with and without situation descriptions) and conducted measurement invariance analyses. Results echoed those of Study 1: Split-half reliabilities of the applied SJTs were low (all estimates below .30), which—considering the small number of items per SJT in Study 2—is in line with the SJT literature (e.g., O'Neill et al., 2012). No significant differences occurred in terms of internal consistency reliabilities between the two SJT versions. Although there was only weak evidence for measurement invariance (invariance of variances and covariances, but no evidence for factorial invariance; see Appendix C), subsequent partial invariance analyses revealed that the loadings of only two items were noninvariant across groups. Similar to Study 1, some fit indices (CFI) were low. In light of these findings and the fact that Hypothesis 2 was formulated at the item level, we again conducted all analyses at the SJT item level.

Hypothesis tests. Hypothesis 2 stated that the construct domain would moderate the difference in SJT item scores when SJT items were administered either with or without situation descriptions. To test Hypothesis 2, we used the same strategy as in Study 1 and conducted a 2 (SJT version: with vs. without situation description) $\times$ 2 (SJT domain: job knowledge and skills vs. other domain) ANOVA with the proportion of correct item responses (separately for all participants in each group) as the dependent variable. This ANOVA did not yield significant main effects for SJT version, $F(1, 56) = 2.68$, $p = .107$, partial $\eta^2 = .003$, or for SJT domain, $F(1, 56) = 0.52$, $p = .821$, partial $\eta^2 = .001$, or for their interaction, $F(1, 56) = 0.67$, $p = .796$, partial $\eta^2 = .001$. Thus, Hypothesis 2 was not supported. Similar to Study 1, we next conducted independent t tests per item and compared the results of these t tests across the three SJT domains. Uncorrected and Bonferroni corrected results ($p < .05/30 = .00167$) are reported in Table 2. Across the three SJTs, 17 out of 30 items (56.7%; when uncorrected p values are used) were identified, resulting in significantly more correct answers (i.e., these items were easier to solve) when situation descriptions were given. After correcting p values, this number dropped to 11 out of 30 items (36.7%) that were identified as resulting in significantly more correct answers.

Table 2
Itemwise Comparison of the Number of Correct Answers Depending on the Availability of Situation Descriptions (Study 2)

Item	Word count of situation description	Glass δ	*t*	*df*	*p*
Pilot Judgment Inventory					
1 (1)	71	−0.20	−2.84	546.29	.003
2 (4)	72	0.29	3.75	547.50	*
3 (14)	99	−0.17	−1.82	544.54	.035
4 (23)	45	−0.03	−0.33	550	.372
5 (25)	71	0.30	3.75	539.78	*
6 (32)	53	0.75	8.64	550	*
7 (33)	53	0.03	0.40	550	.347
8 (40)	36	0.00	−0.12	550	.454
9 (43)	41	−0.20	−2.56	549.86	.006
10 (44)	37	1.42	12.71	429.79	*
Teamwork–KSA Test					
1 (6)	14	−0.32	−3.93	544.79	.000[a]
2 (7)	19	0.07	0.78	550	.217
3 (10)	30	0.96	8.74	430.2	*
4 (11)	20	0.36	4.39	550	*
5 (12)	53	1.14	16.53	483.92	*
6 (14)	34	0.46	6.00	548.58	*
7 (19)	26	0.24	2.47	512.93	.007
8 (26)	31	−0.44	−5.44	549.16	.000[a]
9 (27)	17	0.06	0.70	550	.243
10 (30)	21	−0.02	−0.28	550	.391
Employee Integrity SJT					
1 (4)	54	0.12	1.19	525.28	.117
2 (7)	31	−0.03	−0.59	550	.277
3 (8)	66	0.30	1.95	374.606	.026
4 (10)	61	1.01	12.34	548.315	*
5 (13)	45	0.89	10.44	550	*
6 (14)	47	0.15	1.59	530.756	.057
7 (16)	35	0.46	4.61	463.88	*
8 (18)	101	−0.06	−0.80	550	.212
9 (19)	59	0.14	1.67	550	.048
10 (20)	42	0.10	1.23	533.007	.110

Note. Numbers in parentheses refer to item numbers as used in the original versions of the situational judgment tests (SJTs). One-sided *t* tests. Degrees of freedom vary due to the use of *t* tests for homogeneous and heterogeneous variances. Higher effect sizes reflect more correct answers on items with situation descriptions compared with items without situation descriptions. KSA = knowledge, skills, and ability.
[a] Effect in other than the hypothesized direction and thus not classified as significant.
* $p < .00167$ (p level adjusted to account for alpha inflation).

Visual inspection of these results showed that the number of items that produced significant differences when answered either with or without situation descriptions were higher in the pilot judgment SJT but only when alpha level was not adjusted. That is, among both the teamwork and the integrity SJT items, five out of 10 items differed significantly. Among the pilot judgment SJT items, seven out of 10 items yielded significant differences. After Bonferroni correction, the number of significant items dropped to four (pilot judgment SJT and Teamwork–KSA test) and three (integrity SJT) out of 10 items. The average item effect size was 0.26 (range from −0.44 to 1.42).

When we translated these significant item-level differences into practical score effects, there was a mean difference of about 3 points (on an overall achievable score of 30 points) between the two SJT versions (with and without situation descriptions). When test takers received no situation descriptions, they were able to solve about 17 out of 30 items ($M = 16.52$, $SD = 2.77$, range from 3 to 23) correctly; when test takers received situation descriptions, they were able to solve about 20 out of 30 items correctly ($M = 19.68$ $SD = 2.90$, range from nine to 27).

Additional analyses. To examine the influence of experience on our results, we included experience as a factor (coded as either high or low, that is, less than 10 or more than 20 years of experience as a pilot) in the aforementioned ANOVA on the 30 items as the unit of observation. No main effect for experience, $F(1, 114) = 0.37$, $p = .543$, partial $\eta^2 = .003$, or any interaction with the SJT version factor, $F(1, 114) = 0.34$, $p = .563$, partial $\eta^2 = .002$, or the SJT domain factor, $F(1, 114) = 0.24$, $p = .629$, partial $\eta^2 = .003$, was observed.

Next, we scrutinized the item characteristics of the 11 items that showed significant differences between the two conditions (with and without situation descriptions). Neither item/response option length nor item difficulty was related to differences between the two SJT versions (e.g., correlations between effect sizes in Table 2 and word count of situation description: $r = -.02$). Similar to Study 1, four experts (teamwork experts and pilots) who were

blind to the results of the current study coded the items of the three SJTs as to whether the participants' responses mentioned general versus more context-specific rules of action (interrater agreement of $\kappa = .64$). The differences between items with and without situation descriptions were entered as dependent variable into an ANOVA with SJT domain (job knowledge and skills vs. other domain) and the expert coding (general vs. more context-specific rules of action) as independent factors. The expert coding factor just failed to reach the conventional significance threshold, $F(1, 26) = 3.64$, $p = .067$, partial $\eta^2 = .048$. A significant interaction between the expert coding factor and the SJT domain factor, $F(1, 26) = 4.28$, $p < .05$, partial $\eta^2 = .141$, indicated that SJT items without situation descriptions were more difficult to solve than their counterparts with situation descriptions but only when the SJT item addressed job knowledge and skills and when, at the same time, their correct solution tapped into context-specific rules of action.

Discussion

Study 2 showed that there was no significant interaction between construct domain and SJT version (with vs. without situation description). Across SJT items, it did not make a significant difference whether situation descriptions were included for between 43% (uncorrected alpha level) and 63% (when correcting the alpha level) of the items. In light of results of Study 1, these findings are not surprising for SJT items capturing the content domains of applied social skills and basic personality tendencies. However, we had not anticipated that this result would generalize to job knowledge and skills SJT items assessing pilot judgment in a sample of pilots. Thus, Study 2 shows that the construct domain is less of a driver behind the context-(in)dependent nature of SJT performance than expected.

Some other results of Study 2 are also noteworthy. First, when we inspect the uncorrected results on the item level, there is a slight trend suggesting that there are fewer items from the domain of job-specific knowledge and skills that can be solved without contextual information (30% of the items) than from the other construct domains (50% of the items). Second, there was again a trend (albeit not significant) that response options of items producing significant differences between SJT versions referred to context-specific courses of action. Notably, when those two aspects were combined, a significant interaction revealed that differences were the largest in items addressing job knowledge and skills *and* referring to context specific courses of action in their response options. This latter result qualifies the effect posited in Hypothesis 2. Hence, Study 2 pointed to initial moderators as to when SJT items trigger either context-dependent or context-independent knowledge.

More generally, in both Studies 1 and 2, the presentation of the response options alone seemed sufficient for test takers to successfully identify correct answers for a sizable percentage of the items. How can this be explained? This result can be understood in terms of the two levels of SJT responding that were recently distinguished by Leeds (2012). That is, one must first be sensitive to and detect the subtle differences between the various response options (primary level of processing) before further processing can take place (secondary level of processing). In this primary level of SJT decision making, Leeds argued that test takers evaluate each response option in an absolute ("How effective is this option?"; "Does it make sense?") as well as in a relative sense ("Is this option better than that other option?"). In other words, each of these response options in itself represents a piece of knowledge (a course of action) that test takers compare with their existing knowledge base and with other pieces of knowledge. Only in the secondary decision-making process, test takers then take the situation description (item stem) into account to endorse the prevailing response option.

In order to shed more light on this primary process of evaluating response options, we conducted a third study in which we analyzed test-takers' verbal protocols while responding to the team knowledge SJT without situation descriptions. Accordingly, we examined which strategies and knowledge test takers use when solving SJT items without situation descriptions.

Study 3

Study 3 focused on identifying the different types of context-independent knowledge that test takers use when completing SJT items without situation descriptions. To build hypotheses about this, we took the response option evaluation processes that Leeds (2012) recently proposed as point of departure. According to Leeds, in the primary level of SJT decision making, test takers evaluate the response options in an absolute sense. So, when situation descriptions are not provided, it is likely that test takers evaluate each response alternative according to their general domain knowledge (Motowidlo & Beier, 2010; Motowidlo et al., 2006). This means that they might ask themselves, "Does this response alternative generally make sense?" In the context of team knowledge, for instance, test takers might consider behaviors (e.g., monitoring success or solving problems, motivating team members, being fair, or supporting social climate) that are often included in team effectiveness models (e.g., Morgeson, DeRue, & Karam, 2010) and that are insightful as they list behaviors and courses of action that are generally assumed to have a positive impact on team performance. Hence, we expected that response options that included these general behaviors may be chosen by test takers when the situation description is absent. Thus, we hypothesized the following:

> *Hypothesis 3:* For SJT items without situation descriptions, strategies that refer to behavior regarded as generally positive for team performance will result in more correct responses compared with other or no such strategies.

Besides evaluating response options in an absolute sense, Leeds (2012) further proposed that test takers compare the response options in a relative sense to each other. In fact, in the cognitive ability domain, similar studies have been conducted that revealed that test takers are able to deduce information by comparing all response options with each other (Mittring & Rost, 2008; White & Zammarelli, 1981). Hence, an additional test taker approach for gaining insights from response options in the absence of situation descriptions in SJTs might be to compare response options. Hence, we posited the following:

> *Hypothesis 4:* For SJT items without situation descriptions, strategies that are based on comparing response alternatives will result in more correct responses compared with strategies that are not based on comparing response alternatives.

Method

Participants. A think-aloud technique (concurrent verbal protocol study) was used with a sample of 40 individuals (33% male) who were contacted via postings and mailings. The sample comprised students (63%) as well as employees (37%). The sample of working people included those whose jobs did (21%) or did not (79%) involve leadership responsibilities. All participants reported that their current (student) jobs required working in a team. Their mean age was $M = 25.3$ years ($SD = 4.98$, range = 18–39). Study participation was voluntary.

Measures. In Study 3, we administered again the Teamwork–KSA Test (Stevens & Campion, 1996). Each participant completed either Items 1–18 or Items 19–35 of the Teamwork–KSA Test. We administered only half (Items 1–18 or Items 19–35) of the Teamwork–KSA Test items per each participant to keep test duration at an acceptable level while still enabling participants to elaborate (think aloud) on each item response in detail.

Procedure. The think-aloud technique was employed while participants worked on SJT items. In a 1:1 setting, a trained interviewer asked the participant to read a single SJT item and then articulate any thoughts that came to mind while deciding how to respond to the SJT item (see van Someren, Barnard, & Sandberg, 1994). SJT items consisted of a knowledge-based instruction ("Which of the following would likely be the most effective way to resolve the situation?") and a set of response options but no situation description. The interviewer recorded the participant's explanation and his or her response to this item and then asked the participant to proceed to the next SJT item.

Coding procedure. Two teams consisting of two research assistants each, who were blind to the purpose, design, and hypotheses of the study, coded the responses. The first team developed a coding scheme. They were individually asked to build clusters of similar response strategies. They were then asked to work together and review their clusters, discuss discrepancies, and, if necessary, modify clusters and their definitions. The following categories reflecting general domain knowledge emerged: feasibility of the behavior in applied settings, effectiveness of the behavior for teamwork, general fairness of the behavior due to working in teams, motivating effect of the behavior in the context of teamwork, and the strategy of comparing response options (including expressions with comparative statements). Note that some participants also reported no specific or other strategies (e.g., "Simply sounds like a good solution"). Example statements per category are presented in Table 3.

The second team received the labels and definitions of the clusters obtained after Step 1. They used the following coding protocol: (a) they scanned each response for the response strategies expressed, (b) went through the definitions of each predefined class of strategies, and (c) decided whether the response at hand fell into that category. Coders were given the category definitions and sample responses per category and were free to classify participants' responses into more than one category. Both coders worked independently of each other.

Agreement between the two coders as assessed via Cohen's κ (Cohen, 1960) ranged from moderate to very good (Cohen's κ from .54 to .85; see Table 4). Overall, coders agreed in 89.9% of the coding. In a final step, coders were asked to discuss disagreements and to unify their scores. These unified scores were used for further analyses.

Results

Across all participants and SJT items, the think-aloud technique elicited 665 statements. Participants' statements most frequently fell into the category of "comparison of response options" (44.4%), followed by statements classified as the "effectiveness of the behavior for teamwork" (40.2%). Among all statements, 34.6% were assigned to more than one category (for other frequencies, see Table 4).

Table 3
Example Statements in Verbal Protocol Analysis (Study 3)

Category	Statement (shortened examples)
1. Feasibility of the behavior in question	This solution sounds plausible and applicable in daily routine
	Sounds doable within the typically limited amount of time
	Reflects what is done in real life
2. Effectiveness of the behavior in question	This makes the work process very effective.
	Highly effective since many methods are applied to achieve team goals
	This is a good way to work under time pressure
3. Fairness of the behavior in question	Every opinion should be valued
	Same rights and duties for everybody
	Involving the whole team signals that everybody has a voice in decisions
	A fair distribution of work ensures a good climate
4. Motivating effect of the behavior in question	To ensure high motivation of the team members
	To avoid a demotivating effect of too much work load
	Appreciativeness increases motivation
	This will increase everybody's motivation
5. Comparison of response options	This is the only response option that makes sense
	Among the alternatives, this is the best way to . . .
	This is the only alternative that . . . .

Note. Categories are not assumed to be exhaustive. Coders could assign statements to multiple categories.

Table 4
2 (Correct vs. False Answer) × 2 (Strategy Used vs. Not Used) Cross-Tables

Strategy (% of all statements falling in that category)	SJT item solution		Cohen's κ[a]	Chi-square scores[b]
	Correct	False		
1. Feasibility of the behavior (8.3%)				
Yes	35	37		
No	300	290	.56	0.13, $p = .720$
2. Effectiveness of the behavior (40.2%)				
Yes	167	97		
No	168	230	.54	28.13, $p < .001$
3. Fairness of the behavior (13.7%)				
Yes	59	43		
No	276	284	.76	2.53, $p = .112$
4. Motivating effect of the behavior (10.7%)				
Yes	47	26		
No	288	301	.57	6.23, $p < .05$
5. Use of general knowledge (i.e., use of any one of Strategies 1 to 4)				
Yes	241	163		
No	94	164	.57	33.96, $p < .001$
6. Comparison of response options (44.4 %)				
Yes	159	133		
No	176	194	.85	3.10, $p = .079$

Note. Some statements were classified in multiple categories as they, for example, referred to domain knowledge and a comparison of response options ("Among all alternatives, this is the best way to ensure team success"). SJT = situational judgment test.
[a] Cohen's κ indicates the agreement of the two coders in assigning participants' statements to the categories. [b] Degree of freedom was 1.

Hypotheses 3 and 4 were tested with chi-square tests comparing the frequency of correct responses and false responses in SJT items depending on whether a specific category was used or not. The cell frequencies are presented in Table 4. First, we examined the contingency between SJT item solutions and the use of general knowledge (i.e., focusing on behaviors considered generally positive for team performance). Therefore, we classified verbal expressions that indicated the categories related to feasibility, effectiveness, fairness, or motivating effects into a new broader category, which we labeled the "use of general knowledge" (Table 4). A chi-square test on the basis of this new classification revealed that the use of any of those categories was indeed contingent on SJT item solution, $\chi^2(1) = 33.69$, $p < .001$. Illustrating this result, the use of general knowledge resulted in 1.6 times higher chances of choosing the most effective response option. Thus, Hypothesis 3 was supported by our data.

To further test Hypothesis 3 and to identify single categories that were particularly helpful in identifying the correct answer, we conducted analyses separately for each category. These analyses revealed that the category related to evaluating the general effectiveness of the behavior displayed in response options was contingent on SJT item solutions, $\chi^2(1) = 28.13$, $p < .001$. When this category was used, the ratio of correct versus incorrect responses was about 1.7 (i.e., chances of giving the correct answer were 1.7 times higher than chances of giving an incorrect answer). When this category was not used, the ratio dropped to about 0.7. In other words, chances of giving an incorrect answer were about 1.3 times higher than chances of giving the correct answer. A smaller but also significant effect was found for the category related to evaluating the motivating effect of the behavior listed in the response options, $\chi^2(1) = 6.23$, $p < .05$. However, this category was rarely used (only 73 references were made to this category), and thus results should be interpreted with caution.

Regarding Hypothesis 4, a chi-square test evaluating the contingency between the category related to comparing response options and the SJT item solutions yielded insignificant results, $\chi^2(1) = 2.98$, $p = .084$: The ratio of correct and incorrect answers was almost similar when comparing or not comparing response options (54.4% vs. 47.3%), thereby not supporting Hypothesis 4.[2]

Discussion

Analyzing test-takers' verbal protocols when solving SJT items without situation descriptions provided insights into underlying thought processes that complemented our findings from Studies 1 and 2. In line with our assumptions (see also Leeds, 2012), SJT test-takers used the set of response options as a source of information and, in the absence of situation descriptions, relied on a variety of absolute (general domain knowledge) and comparative strategies for drawing inferences regarding the correct answer. A key finding was that the category of evaluating the general effectiveness of the response in an absolute sense emerged as the only one that significantly helped participants in choosing the correct answer when situation descriptions are lacking. This general effectiveness category referred to the estimation of whether specific responses were most likely to be beneficial for team performance

[2] Ancillary analyses showed that testing Hypotheses 3 and 4 in our sample separately for students and employees yielded similar findings. The only two strategies that were contingent on the frequency of correct responses were the strategies that involved the use of general knowledge and the effectiveness of the behavior in question.

in general (i.e., across most situations). As such, general effectiveness evaluations can be considered "general domain knowledge" (Motowidlo & Beier, 2010). Further research needs to address whether the poor results for the other categories are due to this specific SJT and/or to the situations presented.

General Discussion

Key Conclusions

By comparing results from SJTs with or without situation descriptions, our first objective in this article was to investigate one of the core assumptions underlying SJTs: "How situational is judgment in SJTs?" Across the studies, results showed that for between 43% and 71% of the items, it did not matter whether situation descriptions were included. So, the SJT item stem was an "incidental" feature for at least 43% of the items. As a second objective, we examined various moderators for the extent to which SJT items reflect context-(in)dependent knowledge. We found evidence suggestive of two moderators, namely, the type of construct (job-specific versus generic) and the response option contextualization. These results have at least the following four key implications for SJT practice, research, and theory.

A clearer distinction between the context-dependent and context-independent perspectives in practice and research. Contextualization in SJTs is "expensive" because different groups of subject matter experts are typically consulted for generating contextualized item stems and response options. So far, however, SJT designers and researchers have been left in the dark regarding the extent to which high levels of contextualization (of item stems and response options) are needed. Given this lack of clarity, it is not surprising that there currently exists a lot of variation in the degree to which SJT items (item stems and response options) require more general-domain versus more context-dependent knowledge. Whereas some SJT items use generic item stems and response options that are applicable across a wide variety of situations, others employ highly contextualized item stems and response options. The typical SJT then consists of a mixture of those two item types (as this study's results attest).

Generally, we argue against developing an SJT that contains a heterogeneous mixture of items that tap into both context-dependent and context-independent knowledge. Thus, we posit that making a sharper distinction between the context-dependent and context-independent perspectives might advance the SJT field because these different perspectives have different conceptualizations, design implications, and purposes. Although being speculative and awaiting further empirical support, we suggest that when SJTs are developed for or are applied in entry-level selection, admissions, and recruitment purposes, high levels of contextualization are not needed, and it might suffice to use short generic item stems that approximate job-related situations (e.g., "A customer is rude to you; what do you do?"). Such generic SJT item stems might be followed by response options that are applicable across a wider range of situations and tap into general domain knowledge (implicit trait policies; Motowidlo & Beier, 2010; Motowidlo et al., 2006). At a practical level, substantial time and cost savings can be made when developing SJTs for such purposes as various groups of subject matter experts might not be needed for developing them. The SJT content development process can be even further streamlined by removing the item stems. In such single-response SJTs, only item options (with a minor level of contextualization) are presented to test takers (e.g., Crook et al., 2011). Again in absence of empirical support, one may speculate that one might continue designing SJT items with high levels of contextualization when SJTs are developed for advanced-level selection, specialized training, and certification purposes. In those cases, applicants should have already acquired the requisite fine-grained procedural and declarative knowledge so that the highly contextualized SJT item stems and response options zoom into this context-dependent knowledge.

A research agenda on contextualization. Our examination of the extent to which SJT performance depends on context-dependent versus context-independent knowledge constitutes an important first step to put contextualization on the agenda of SJT researchers. We see three broad avenues as critical for future research on contextualization in SJTs and other selection procedures. A first avenue of research consists of contrasting the criterion-related validity of SJTs without situation descriptions to the one of traditional SJTs (with situation descriptions). One expectation is that the removal of situations reduces the point-to-point correspondence with the criterion and therefore impairs predictive accuracy. However, another expectation might be that SJTs without situation descriptions make the SJT more ambiguous and ill-defined as there are fewer cues available. In turn, this might lead to a higher correlation with cognitive ability (as more cognitive resources are required to come up with a solution) and a higher correlation with social skills (as one needs the ability to infer the situational cues). Consistent with research on ambiguous demands in assessment centers (Jansen et al., 2013), SJTs without situation descriptions might therefore produce good criterion-related validities especially for jobs with ill-defined demands.[3]

Second, it is important to examine to what extent the positive applicant reactions toward SJTs are connected to their level of contextualization. That is, we do not know whether applicants react more favorably toward contextualized versus more generic SJT items. In the personality field, that does not seem to be the case (Holtz, Ployhart, & Dominguez, 2005). Perhaps the SJT purpose and type of job serve as moderators. Similarly, we do not know whether applicants equate contextualization with job relatedness and how it relates to the potential of SJTs as realistic job previews. So, future research is needed to examine effects of SJT contextualization on applicant perceptions.

Third, we need to scrutinize the effects of contextualization on construct measurement in SJTs. Keeping the contextualization of all items at the same level in an SJT (with the level being contingent on the criterion specificity to be predicted) should remove some of the item heterogeneity that has been posited to lead to poor measurement models in SJTs. Generally, we posit that continued efforts should be undertaken to improve con-

[3] As an initial investigation of criterion-related validity, we examined correlations of the pilot SJT versions (with and without situation description) with flight experience indicators. Correlations of the SJT scores with these criteria were low (*r*s from .02 to .19) and similar to cognitive predictors of work experience (Dragoni, Oh, Vankatwyk, & Tesluk, 2011). Notably, neither version produced SJT scores with significantly different correlations with experience (*Z* scores 1.24 and 1.45, *p*s $>$ .05).

struct measurement in SJTs (see the construct-driven SJT approach of Ployhart, Porr, & Ryan, 2004, for an example).

The importance of reexamining assumptions underlying SJTs. A third main implication for research on SJTs is that we may have been somewhat naïve in assuming that inserting situational cues in assessments automatically would allow them to tap into context-dependent knowledge. Our results show that the perception of the situation does not matter in about half of the items, which runs counter the interactionist underpinnings of SJTs. So, conceptually, this study's results might serve as an eye opener to SJT designers and researchers, prompting them to reexamine their assumptions. One example is that currently SJTs typically require people to indicate how to respond to the situation. It is striking that questions about the situation itself are not explicitly included in extant SJTs. However, it should be possible to precede the typical SJT question about one's reactions to the situation with a question about one's perception of the situation (e.g., What is the situation about? What are the thoughts of the persons involved in the situation? see Rockstuhl, Ang, Lievens, & Van Dyne, in press). Adding a situation-perception question to the traditional situation-response question might put not only more emphasis on the situation component of SJTs, but it might also ground SJTs more firmly in the interactionist paradigm. As outlined in the seminal work of Endler and Magnusson (1976), an interactionist approach inherently takes into account both the person's perception of situations (i.e., a stimulus-analytical approach) and his or her reaction to these situations (i.e., response-analytical approach). So far, SJT design and research have mainly adopted the response-analytical approach (Campion & Ployhart, 2013).

Another example is that an interactionist (person $\times$ situation) framework is seldom considered explicitly in developing SJTs as they are typically based on a competency model or critical incidents. Along these lines, Campion and Ployhart (2013) referred to the taxonomy of Murtha, Kanfer, and Ackerman (1996) in which traits (divided into subfactors) and situations (sorted according to their psychological meanings) are fully crossed to form a situational-behavioral taxonomy. A similar strategy consists of developing SJT items that build "If . . . , then . . ." contingencies (Mischel & Shoda, 1995) into the items. Using such strategies might lead to SJT design and research becoming truly interactionist.

Toward a theory of the context-(in)dependency of SJT performance. A final contribution is that we could use this study's results to start building a theory that delineates which factors determine whether SJTs assess context-dependent versus context-independent knowledge. In line with the construct-method distinction, we distinguish between two main moderators: (a) construct and (b) measurement method moderators.

On the construct level, the nature of the underlying construct is the main factor because some constructs are by nature more specific (job knowledge and skills) than others (applied social skills and basic personality tendencies). However, results of Study 2 show that this will not automatically lead to SJT performance being determined by context-(in)dependent knowledge. That is because how the construct is operationalized and how the construct is measured via the SJT methodology also seem to matter. The notion that the nature of the construct does not automatically lead to SJTs performance being determined by context-(in)dependent knowledge is consistent with knowledge acquisition models and research demonstrating that more specific and context-dependent knowledge typically accrues from general domain knowledge. Motowidlo and Beier (2010) exemplified this as follows for a generic construct (basic personality tendency) such as agreeableness:

> Someone can accurately believe, and therefore "know," that agreeable action is generally more effective than disagreeable action but then learn exceptions to the rule in his or her particular job and discover that the most effective action in some specific situations expresses disagreeableness and that in other specific situations agreeableness is simply irrelevant for effectiveness. (p. 323)

This quote exemplifies that even a basic personality tendency might be operationalized with context-specific items and response options, whereas it is in principle equally possible to operationalize even a context-specific construct, such as, for instance, pilots' job knowledge with general item scenarios and item options reflecting general rules of action.

Thus, an important issue seems to be whether a construct is measured via SJT items in such a way that these items tap into more fine-grained contextualized knowledge about the constructs. Along these lines, our theory posits that such measurement moderators of the context dependency of SJTs include the level of contextualization of the item stem and the response options. The item stem presents the context to test takers, thereby activating the knowledge constructs that might be needed to answer the situation. Some item stems might be constructed as highly contextualized, whereas others might be rather generic. Highly contextualized item stems typically activate context-dependent knowledge in test takers (as long as they possess it). In turn, SJT response options represent various ways of responding to the situation depicted in the item stem. Accordingly, they operationalize the behavioral domain represented by the construct(s) assessed by the SJT. In accordance with the item stems, response options vary in their level of contextualization. Some reflect broadly applicable courses of action (e.g., in team work settings) that capture general domain knowledge. Conversely, response options representing more contextualized courses of action require more fine-grained and specific context-dependent knowledge. This study provided initial evidence, suggestive that when response options reflect more context-specific courses of action, SJT performance becomes more context dependent.

Taken together, this theory predicts that assessing context-dependent knowledge with SJT items is most likely to occur when job knowledge and skill constructs are assessed with highly contextualized item stems and response options. Conversely, SJTs most likely tap into context-independent knowledge when broad constructs are assessed with generic item stems and generic response options.

Future research is needed to test the relative importance of the different factors of this theory. To this end, SJT items that cross these different construct and method factors with each other can be specifically developed (e.g., the same item can be made more or less contextualized). In our research, existing SJTs were used for external validity reasons, precluding such a controlled examination of the relative importance of each of these factors. We also believe that this construct-method theory for conceptualizing contextualization might prove useful for

other sample-based predictors such as situational interviews, work samples, or assessment centers.

Limitations

Among the limitations of the reported studies is that our conclusions and implications are based on only three different SJTs. So, one may speculate about the generalizability of our findings to other SJTs. For example, one might question whether our results apply to more realistic SJT presentation formats (e.g., video). It might well be that the situation gains in importance when more realistic stimulus formats are used. The context in these formats is multilayered, including verbal, nonverbal, and paralingual cues. So, we encourage replications with context-richer SJT formats. Conceptually, such studies are informative as to which SJT features represent "radicals" instead of "incidentals."

Another generalizability aspect relates to the sample. The relevance of situation descriptions in SJTs may be contingent on the cultural background of individuals, insofar as some cultures may be better able "to fill in the blanks" (the missing situation descriptions) than other cultures. For instance, cross-cultural research has demonstrated that cultures may either be high or low on context (Gudykunst et al., 1996; Hall, 1976; Kittler, Rygl, & Mackinnon, 2011). Similarly, individuals from either universalistic or particularistic cultures may show differences in their tendency to apply general domain knowledge. Perhaps an investigation of subgroup differences in SJTs without situation descriptions provides an interesting new twist to the adversity-validity dilemma.

Finally, we scrutinized test takers' verbal protocols only in the condition without situation descriptions. Although we expect that in traditional SJTs, similar strategies might also underlie their thought processes (see Leeds, 2012), in the future, researchers should compare verbal protocols in SJTs with and without situation descriptions. Eye movement analysis might also elucidate how much time people spend on reading the situation descriptions compared with the response options and how frequently they backtrack to these situation descriptions.

Conclusion

As "situational" is an explicit part of the term SJT, it has been traditionally assumed that test takers require a contextual description for solving SJT items. This study did not take this assumption for granted and sought to examine how "situational" judgment on SJT items actually is. We found that even without a contextual description test takers could solve on average more than half of the items of various existing SJTs. This result does not support the traditional contextualized perspective underlying SJTs for a large set of SJT items. There was initial evidence that judgment in SJTs became more situational when (a) items measured job knowledge and skills and (b) response options denoted context-specific rules of action.

Conceptually, these results bring up questions regarding the interactionist assumptions underlying SJTs. At a practical level, they suggest that generic SJT items that approximate the work context might suffice for SJTs developed for entry-level, admissions, and recruitment purposes. Future SJT research should examine when and why contextualization matters for SJT performance, applicant perceptions, and validity.

References

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

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. doi:10.1207/s15328007sem1201_3

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13,* 225–232. doi:10.1111/j.1468-2389.2005.00319.x

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

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling, 11,* 272–300. doi:10.1207/s15328007sem1102_8

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.

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

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. doi:10.1177/001316446002000104

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

Dragoni, L., Oh, I. –S., Vankatwyk, P., & Tesluk, P. E. (2011). Developing executive leaders: The relative contribution of cognitive ability, personality, and the accumulation of work experience in predicting strategic thinking competency. *Personnel Psychology, 64,* 829–864. doi:10.1111/j.1744-6570.2011.01229.x

Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin, 83,* 956–974. doi:10.1037/0033-2909.83.5.956

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

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62,* 451–482. doi:10.1146/annurev-psych-120709-145346

Glass, G. V., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gudykunst, W. B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K. S., & Heyman, S. (1996). The influence of cultural individualism-

collectivism, self construals, and individual values on communication styles across cultures. *Human Communication Research, 22,* 510–543. doi:10.1111/j.1468-2958.1996.tb00377.x

Hall, E. T. (1976). *Beyond culture*. New York, NY: Doubleday.

Holtz, B. C., Ployhart, R. E., & Dominguez, A. (2005). Testing the rules of justice. The effects of frame-of-reference and pre-test information on personality test responses and test perceptions. *International Journal of Selection and Assessment, 13,* 75–86. doi:10.1111/j.0965-075X.2005.00301.x

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. doi:10.1080/10705519909540118

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86,* 897–913. doi:10.1037/0021-9010.86.5.897

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

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

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

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouche, M. (2006). From the subjects' point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment, 22,* 168–176. doi:10.1027/1015-5759.22.3.168

Kasten, N., & Freund, P. A. (2013, September). *Eine metaanalytische Reliabilitätsgeneralisierung von Situational Judgment Tests (SJTs) [A meta-analytic reliability generalization of situational judgment tests]*. Paper presented at the 12th meeting of the section Differential Psychology, Personality Psychology, and Psychological Assessment of the German Psychological Society, Greifswald, Germany.

Kittler, M. G., Rygl, D., & Mackinnon, A. (2011). Special review article: Beyond culture or beyond control? Reviewing the use of Hall's high-/low-context concept. *International Journal of Cross-Cultural Management, 11,* 63–82. doi:10.1177/1470595811398797

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

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. doi: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., & 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. doi:10.1037/0021-9010.91.5.1181

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92,* 1043–1055. doi:10.1037/0021-9010.92.4.1043

McCarthy, J. M., & Goffin, R. D. (2003). Is the test attitude survey psychometrically sound? *Educational and Psychological Measurement, 63,* 446–464. doi:10.1177/0013164403063003007

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,* 63–91. doi:10.1111/j.1744-6570.2007.00065.x

Mischel, W., & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102,* 246–268. doi:10.1037/0033-295X.102.2.246

Mittring, G., & Rost, D. H. (2008). Die verflixten Distraktoren: Über den Nutzen einer theoretischen Distraktorenanalyse bei Matrizentests [The darn distractors: On the usefulness of theoretical distractor analyses in matrix tests]. *Diagnostica, 54,* 193–201. doi:10.1026/0012-1924.54.4.193

Morgeson, F. P., DeRue, D. S., & Karam, E. P. (2010). Leadership in teams: A functional approach to understanding leadership structures and processes. *Journal of Management, 36,* 5–39. doi:10.1177/0149206309347376

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95,* 321–333. doi:10.1037/a0017975

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

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

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

Motowidlo, S. J., Martin, M. P., & Crook, A. E. (2013). Relations between personality, knowledge, and behavior in professional service encounters. *Journal of Applied Social Psychology, 49,* 1851–1861.

Murtha, T., Kanfer, R., & Ackerman, P. L. (1996). Toward an interactionist taxonomy of personality and situations: An integrative situational-dispositional representation of personality traits. *Journal of Personality and Social Psychology, 71,* 193–207. doi:10.1037/0022-3514.71.1.193

Neisser, U. (1976). General, academic, and artificial intelligence. In L. Resnick (Ed.), *The nature of intelligence* (pp. 135–144). Hillsdale, NJ: Erlbaum.

O'Neill, T., Goffin, R. D., & Gellatly, I. R. (2012). The knowledge, skill, and ability requirements for teamwork: Revisiting the Teamwork-KSA Test's validity. *International Journal of Selection and Assessment, 20,* 36–52. doi:10.1111/j.1468-2389.2012.00578.x

Ployhart, R. E., Porr, W., & Ryan, A. M. (2004, April). *New development in SJTs: Scoring, coaching and incremental validity*. Paper presented at the 19th annual convention of the Society for Industrial and Organizational Psychology, Chicago, IL.

Potosky, D., & Bobko, P. (2004). Selection testing via the Internet: Practical considerations and exploratory empirical findings. *Personnel Psychology, 57,* 1003–1034. doi:10.1111/j.1744-6570.2004.00013.x

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. doi:10.1037/0021-9010.85.6.880

Rockstuhl, R., Ang, S., Lievens, F., & Van Dyne, L. (in press). Putting judging situations into situational judgment tests: Verbal protocols and incremental validity. *Journal of Applied Psychology*.

Stevens, M. J., & Campion, M. A. (1994). The knowledge, skills, and ability requirements for teamwork: Implications for human resource management. *Journal of Management, 20,* 503–530. doi:10.1016/0149-2063(94)90025-6

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

Stevens, M. J., & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management, 25,* 207–228. doi:10.1016/S0149-2063(99)80010-5

Thorndike, R. L., & Stein, S. (1937). An evaluation of the attempts to measure social intelligence. *Psychological Bulletin, 34,* 275–285. doi: 10.1037/h0053850

Thornton, G., & Cleveland, J. (1990). Developing managerial talent through simulation. *American Psychologist, 45,* 190–199. doi:10.1037/0003-066X.45.2.190

Tippins, N. T., & Adler, S. (2011). *Technology-enhanced assessment of talent*. San Francisco, CA: Jossey–Bass. doi:10.1002/9781118256022

Topping, K., Bremner, W., & Holmes, E. A. (2000). *Social competence. The handbook of emotional intelligence*. San Francisco, CA: Jossey–Bass.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London, England: Academic Press.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436–458. doi:10.1037/0022-3514.49.2.436

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.

White, A. P., & Zammarelli, J. E. (1981). Convergence principles: Information in the answer sets of some multiple-choice intelligence tests. *Applied Psychological Measurement, 5,* 21–27. doi:10.1177/014662168100500103

Appendix A

Sample Item of the Teamwork–KSA Test

Sample item of the Teamwork–KSA Test as administered in the SJT version with situation descriptions (Stevens & Campion, 1994):

> Suppose that you find yourself in an argument with several co-workers about who should do a very disagreeable but routine task. Which of the following would be the most effective way to resolve this situation?
>
> (a) Have your supervisor decide, because this would avoid any personal bias.
>
> (b) Arrange for a rotating schedule so everyone shares the chore. (*correct answer*).
>
> (c) Let the workers who show up earliest choose on a first-come, first-served basis.
>
> (d) Randomly assign a person to do the task and don't change it.

Sample item of the Teamwork–KSA Test as administered in the SJT version without situation descriptions follows. Please note that participants were informed at the beginning that the test was about teamwork and were given the following global instruction: "Decide for each item which one of the response options would be most effective or reflect an ideal behavior."

> (a) Have your supervisor decide, because this would avoid any personal bias.
>
> (b) Arrange for a rotating schedule so everyone shares the chore.
>
> (c) Let the workers who show up earliest choose on a first-come, first-served basis.
>
> (d) Randomly assign a person to do the task and don't change it.

(*Appendices continue*)

Appendix B

Goodness-of-Fit Statistics and Tests of Invariance for the Items of the Teamwork–KSA Test (Study 1)

Model description	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	RMSEA [90% CI]	SRMR
Single group models (with situation description)							
Single-factor model	699.32**	560			.53	.035 [.025, .042]	.068
Two-factor model (interpersonal & self-management)	694.29**	559			.54	.034 [.025, .042]	.068
Five-factor model (five factors of the Teamwork–KSA Test)	—[a]	—			—	—	—
Single group models (without situation description)							
Single-factor model	707.75**	560			.33	.038 [.029, 046]	.073
Two-factor model (interpersonal & self-management)	692.46**	559			.40	.036 [.026, .045]	.073
Five-factor model (five factors of the Teamwork–KSA Test)	—[a]	—			—	—	—
Multigroup models[b]							
Baseline model: two factor model (unconstrained)	1386.75**	1118			.48	.025 [.020, .029]	.068
Only factor loadings constrained	1433.52**	1151	46.77	33	.45	.025 [.021, .029]	.071
Only variance and covariance constrained	1390.10**	1121	3.35	3	.48	.025 [.020, .029]	.069
Model fully constrained to equality	1437.77**	1154	51.02*	36	.45	.025 [.021, .029]	.072

Note. Sample sizes in the two groups were $n = 210$ (with situation description) and 185 (without situation description); we excluded participants with missing values; missing values were completely at random as indicated by an insignificant MCAR (missing completely at random) test: $\chi^2(982) = 967.43$, $p = .624$. The two-factor model was chosen due to its significantly lower chi-square value in both conditions. Please note that O'Neill et al. (2012) obtained an equally low value for the relative noncentrality index (RNI), which is equivalent to comparative fit index (CFI). KSA = knowledge, skills, and ability; RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardized root-mean residual.
[a] Model could not be identified. [b] Groups: with situation description and without situation description.
* $p < .05$. ** $p < .01$.

Appendix C

Goodness-of-Fit Statistics and Tests of Invariance for the Situational Judgment Tests Items (Study 2)

Model description	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	RMSEA [90% CI]	SRMR
Single group models (with situation description)							
Single-factor model	589.91**	405			.33	.039 [.032 .046]	.063
Two-factor model (aviation and combined teamwork + integrity factor)	581.22**	404			.36	.038 [.031 .045]	.063
Three-factor model (aviation, teamwork, integrity)	—[a]	—			—	—	—
Single group models (without situation description)							
Single-factor model	482.56**	405			.47	.028 [.016 .037]	.060
Two-factor model (aviation and combined teamwork + integrity factor)	480.62**	404			.48	.027 [.016 .036]	.060
Three-factor model (aviation, teamwork, integrity)	—[a]	—			—	—	—
Multigroup models[b]							
Baseline model: two factor model (unconstrained)	1061.83**	808			.40	.024 [.020 .028]	.063
Only factor loadings constrained	1109.33**	836	47.50*	28	.35	.024 [.020 .028]	.065
Only variance and covariance constrained	1064.12**	811	2.29	3	.40	.024 [.020 .028]	.063
Model fully constrained to equality	1111.75**	839	49.92*	31	.35	.024 [.020 .028]	.065
Partial invariance model[c]	1098.15**	834	11.18*	2[d]	.37	.024 [.020 .028]	.064

Note. Sample sizes in the two groups were $n = 298$ (with situation description) and 254 (without situation description). Please note that O'Neill et al. (2012) obtained an equally low value for the relative noncentrality index (RNI), which is equivalent to comparative fit index (CFI). The two-factor model was chosen due to its significantly lower chi-square value in the condition with situation description. RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardized root-mean residual; KSA = knowledge, skills, and ability.
[a] Model could not be identified. [b] Groups: with situation description and without situation description. [c] Additional partial invariance analyses revealed that factor loadings of only two items (Item 3 on the integrity situational judgment test and Item 6 on the Teamwork–KSA Test) were not invariant across groups; this model's results (with all except two factor loadings constrained to equality) are also presented here. [d] Comparison with the model in which all factor loadings were constrained.
* $p < .05$. ** $p < .01$.

Received April 26, 2013
Revision received July 7, 2014
Accepted July 14, 2014 ■