

Different Culture, Same Situation? Translating and Applying a Situational Judgment Test From Germany in Cuba

Philipp Schäpers¹, Henrik Heinemann¹, Daybel Pañellas Alvarez², Laura Nohr³,
and Franz W. Mönke¹

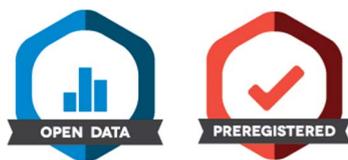
¹Department of Psychology, University of Münster, Germany

²Faculty of Psychology, Universidad de La Habana, Cuba

³Department of Psychology, Freie Universität Berlin, Germany

Abstract: Situational judgment tests (SJTs) are popular instruments in selection and assessment. However, the application of SJTs to non-Western cultural contexts remains scarce. In this study, we investigated whether an SJT on personal initiative, developed in Germany and translated into Cuban Spanish, had similar psychometric properties in Cuba. Second, there is an ongoing debate about the extent to which the situation description plays an important role for SJTs. We supposed that the impact of situation descriptions might depend on test takers' familiarity with the culture in which the SJT was developed. Hence, we tested whether the omission of situation descriptions had larger effects in a Cuban than in a German sample. We applied a 2 (with vs. without situation description in the item stem) \times 2 (cultural background: Cuba vs. Germany) between-subjects design ($N_{\text{Cuba}} = 192$, $N_{\text{Germany}} = 213$). The results revealed similar psychometric properties between Cuban and German test takers concerning measurement invariance, construct-related validity, and reliability. In addition, we examined whether samples differ regarding applicant perceptions: Notably, for four of six applicant perception scales, the Cuban sample reported a more positive view of the SJT. Furthermore, we found that the effect of situation availability on SJT performance did not substantially depend on the test takers' cultural background. Implications for cross-cultural generalizability are discussed.

Keywords: situational judgment test, contextualization, psychometric evaluations



Situational judgment tests (SJTs) are widely used assessment tools that have been applied in various contexts (e.g., medical education and training, personnel selection, or personality assessment; Mussel et al., 2018; Patterson et al., 2016). This test format is commonly conceptualized as a low-fidelity simulation and follows, similar to simulative assessments, the principle of behavioral consistency and point-to-point correspondence. That is, test takers are usually presented with a variety of challenging (text or video) situation descriptions and various response options on how to react to these situations.

Although SJTs are popular in practice and research, most SJTs have been developed in and, thus, might be limited to a specific Western context. Thus, little is known about whether established SJTs can be translated and transferred to non-Western cultures (e.g., Herde et al., 2019), as the interpretation of SJT situations may be contingent on the test takers' cultural background. Furthermore, the context (in)dependence of SJTs has been controversially discussed in recent years (e.g., Freudenstein et al., 2020; Lievens & Motowidlo, 2016): That is, various research found that situation descriptions (i.e., item stems) might play a less important role in response behavior than previously thought (Krumm et al., 2015).

The objectives of this cross-cultural study are twofold: First, we investigated whether an SJT on personal initiative, developed and validated in Germany and

translated into Cuban Spanish, had similar psychometric properties in Cuba. Second, we tested whether the omission of situation descriptions had larger effects in a Cuban sample than in a German sample.

Study Background and Hypotheses

Situational Judgment Tests and Their Cross-Cultural Generalizability

SJTs have been widely applied in various high-stakes contexts of selection and assessment since the use of SJTs results in favorable applicant perceptions (e.g., Kanning et al., 2006) and substantial criterion-related validity (Christian et al., 2010). Furthermore, SJTs are often preferred over self-ratings in high-stakes settings because they are more difficult to fake (Kanning & Kuhne, 2006). As low-fidelity simulations, one might assume that SJTs work in a similar way as other simulative selection procedures (e.g., situational interview questions or role plays). That is, “SJTs provide relevant context to applicants so that they can imagine themselves in a particular scenario and apply their context-dependent knowledge to respond to it” (Lievens et al., 2021, p. 287). Thus, SJTs are based on the principle of behavioral consistency by establishing a direct correspondence between the simulated content (SJT items) and the target construct (e.g., future job performance). Test takers put themselves in the situation presented and indicate how they would or should behave in the presented situation with the help of the response options. Hence, characteristics of the presented situation description might be crucial as it determines how participants react via the presented response options.

However, the application of SJTs to non-Western cultures remains scarce. One reason for this is that SJTs mimic highly contextualized situations, i.e., critical work situations that are related to a very specialized scenario such as certain job demands or company characteristics (Campion & Ployhart, 2013). Thus, applying an SJT in a different culture may lead to different psychometric outcomes: First, the interpretation of the situation descriptions could differ between cultures. Thus, we argue that test takers might focus on other aspects of the situation description; also, certain situations might not be as relevant in another culture. Second, test takers might use different strategies to react to the situations, as interpersonal norms, interpretation of efficient behavior, and relevance of the problem may vary across regions. Consequently, third, the original scoring of the SJT may no longer be adequate and could reflect a different target construct. For instance, various interactions at work differ greatly between different cultures, so one would expect differences in interpretation and behavior here. Thus, changing the cultural context might also change how an SJT works.

Comparing SJT Properties in Different Cultures: Cuba and Germany

Most SJTs have been developed in Western cultures, e.g., the United States and Europe. In our study, we examine how an established work-related SJT, originally developed in Germany, can be transferred to a Caribbean/Latin American culture. The Cuban culture differs from Germany in various (test-related) aspects: First, it is influenced by a history of Spanish colonization, slavery, and the socialistic revolution in the 1950s. In classifications, Cuba is seen as a collectivistic culture (e.g., Díaz Bravo & Pañellas Alvarez, 2019; Galati et al., 2004); i.e., “individuals experience themselves in relation to the social environment” (Nohr et al., 2021, p. 4). Hence, this may result in different perceptions and interpretations of the assessed construct among the participants (see Van de Vijver & Tanzer, 2004). Second, the official language in Cuba is Spanish, and most inhabitants speak Cuban Spanish – a form of Caribbean Spanish influenced by dialects from West Africa and France. Language as one of the most salient aspects of cultural differences might impact test takers on various ways, e.g., on how attention is directed (Fausey et al., 2010; Reali et al., 2006) or what aspects of the item will be memorized (Geisinger, 2003). Third, Cuba has a state-controlled economic system, although recent reforms have allowed some forms of private businesses. In contrast, Germany, with its social market economy, presents a very different culture to employees: Private companies play a much larger role, and former state businesses (e.g., postal and train services) are now commercially driven companies. Also, Germany is seen as an individualistic society, and German is the dominant language. According to Hofstede’s five dimensions of cultural values (i.e., power distance, uncertainty avoidance, individualism, masculinity, and long-term orientation), the Cuban culture is characterized by collectivism, short-term orientation, a high-power distance, and a high uncertainty avoidance (e.g., Banai & Reisel, 2007), whereas Germany can be described as more long-term oriented, with less power distance and a higher focus on individualism (see Hofstede, 2001). Consequently, Cuba and Germany differ in cultural, work-related, and language-related aspects. As SJTs can be conceptualized as a simulative selection procedure that presents specific working situations, one might assume that test takers with another cultural background will react to these situations differently. For instance, (potential) conflict situations with colleagues or customers are a frequent issue in SJT items (see, e.g., SJT Item No. 4 of the SJT on personal initiative; Bledow & Frese, 2009). Expected behaviors and rules of conduct vary significantly between different cultures, so it can be expected that a participant might behave in a way that is common for their culture, but this does not correspond to the solution key that was developed for a

sample with another cultural background. Notably, this may change psychometric properties such as construct-related validity, test performance, or reliability. Hence:

Research Question 1: Do the psychometric properties of an SJT (that was developed and validated in Germany) differ between Germany and Cuba?¹

The Impact of Contextualization on SJTs

We are aware of no reports concerning the use of SJTs in Cuban organizations and enterprises: Cuban participants might not be as familiar with SJTs as a German sample. Furthermore, one might argue that an SJT that was developed for German test takers includes situations that are common for German workplaces that might be less typical in Cuba. Hence, we hypothesize:

Hypothesis 1: Applicants' perceptions of situation typicality, procedural fairness dimensions (face validity, perceived predictive validity, opportunity to perform, and perceived knowledge of results), positive affect (enjoyment), and test-taking motivation will be higher for German test takers than Cuban test takers.

There is an ongoing controversy about the extent to which the situation description plays an important role for the SJT (e.g., Lievens & Motowidlo, 2016). Although the situation description can be seen as the heart of any SJT (e.g., Campion & Ployhart, 2013), suggesting that SJT performance depends on the presented context, Krumm et al. (2015) found that a large proportion of SJT items (between 43% and 71%) could be correctly solved, even when situation descriptions were completely removed. Further studies extended these findings: Key outcomes such as construct-related validity, criterion-related validity, and applicant perceptions were not (or only marginally) affected by removing situation descriptions from SJTs (Schäpers, Mussel, et al., 2020). Even fidelity (video vs. text-based SJTs; Schäpers, Lievens, et al., 2020) or format of the SJT (traditional vs. construct-driven SJTs; Schäpers, Freudenstein, et al., 2020) did not explain the reported findings. However, all these findings were based on Western samples, namely from the United States and Europe. We suppose that this finding might depend on test takers' familiarity with the culture in which the SJT was developed: That is, German participants might be more familiar with an SJT that was developed for a German working context (here, an SJT on personal initiative) than Cuban test takers. Hence, without situation descriptions, German test takers might have

similar scores to their scores on the initial SJT version (with situation descriptions), whereas Cuban test takers could come to different conclusions about what kind of situation is intended by the items. Thus, we hypothesize:

Hypothesis 2: Situation descriptions are more important for Cuban participants to identify the correct answer on an SJT on personal initiative than for a German sample.

Method

Our hypotheses, study design, and analyses were pre-registered; see https://aspredicted.org/9JX_8PL. Data and analysis code are available in the Electronic Supplementary (ES) at the Open Science Framework (<https://bit.ly/3gQRHih>).

Sample

Based on an a priori power analysis, we aimed to assess 360 participants (small-medium effects: $f = 0.175$, $1 - \beta = .80$). During data preparation, we excluded 46 Cuban and 31 German participants due to careless responding (i.e., voluntary self-exclusion and failed instructed response items; Meade & Craig, 2012) and missing data (listwise deletion; 19 Cuban and 26 German test takers). Our final sample consisted of 192 Cuban and 213 German participants.

Cuban participants (convenience sample) were on average 36.2 years old ($SD = 12.3$, range 19–75), 30.3% identified as men, and 69.7% as women. On average, Cuban participants worked 38.6 h a week ($SD = 17.5$, range 0–110) with 12.7 years of job experience ($SD = 12.4$, range 0–50). German participants (Prolific panel) were on average 31.1 years old ($SD = 10.2$, range 18–71), 49.3% identified as men, 48.4% as women, and 2.3% as nonbinary. German participants averaged 28.4 working hours a week ($SD = 16.2$, range 0–105) with 7.3 years of job experience ($SD = 8.5$, range 0–43).

Procedure and Translation

First, following the suggested procedures by Brislin (1970) and Jones et al. (2001), we translated the SJT and measures into Cuban Spanish. For that purpose, two Cuban natives

¹ Following a native speaker's advice, we slightly changed the wording of the preregistered research question RQ1 for improved readability only. Importantly, we have made no changes to the content.

(subject matter experts: professors of psychology) independently translated all measures into Cuban Spanish. Then, a third, bilingual subject matter expert (fourth author of this manuscript), who was blind to the original version, compared these initial Spanish versions and back-translated a reconciled version. We compared the original and back-translated measures; minor variations were solved together with the bilingual expert. A German version of the SJT was provided by Bledow and Frese (2009). Then, we applied a 2 (situation description: with vs. without situation description in the item stem) \times 2 (culture: Cuba vs. Germany) between-subjects design to test our hypotheses. All data were collected online: After obtaining informed consent, we asked participants to answer the SJT on personal initiative. Afterward, we assessed applicant perceptions, self-reported personal initiative, test-taking motivation, and demographics. The data for the Cuban sample were collected in May and June 2022, and the data with German test takers were collected in July 2022. The scientific committee at the Faculty of Psychology of the Universidad de La Habana approved our procedure.

Measures

Personal Initiative: SJT

The 12-item SJT on personal initiative was developed and validated by Bledow and Frese (2009). As suggested by the authors, each SJT item had four or five response options: We asked participants to indicate which of the presented behaviors they would perform *most likely* and *least likely*. If they selected a response option representing high personal initiative as *most likely*, their answer was coded as +1; if they selected an option with a medium

level of personal initiative, 0; and if they chose a response option indicating low personal initiative, -1 (vice versa for *least likely* ratings). The mean of all item scores represented the overall SJT score. Due to a technical admission error in the Cuban sample (repetition of a response option), we had to exclude Item 2 from our analyses, resulting in 11 SJT items.

Personal Initiative: Self-Rating

We assessed a self-rating of personal initiative with the 7-item scale by Frese et al. (1997), e.g., “I actively attack problems” (5-point rating scale: 1 = *strongly disagree*, 5 = *strongly agree*). Reliability was good, with McDonald’s ω values ranging between .76 and .87 (see Table 1).

Applicant Perceptions

We assessed applicant perceptions by the following measures: face validity (5 items, e.g., “I did not understand what the examination had to do with the job”; Smither et al., 1993), perceived predictive validity (5 items, e.g., “I am confident that the examination can predict how well an applicant will perform on the job”; Smither et al., 1993), perceived knowledge of results (3 items, e.g., “After I finished the examination it was clear to me how well I performed”; Smither et al., 1993), positive affect (2 items, e.g., “I enjoyed the examination to a great degree”; Smither et al., 1993), chance to perform (4 general items, e.g., “I could really show my skills and abilities through this test,” four specific items, e.g., “I could really show my skills and abilities regarding personal initiative through this test”; Bauer et al., 2001), and test-taking motivation (5 items, e.g., “I wanted to do well on this test”; Arvey et al., 1990). All scales were rated on 5-point rating scales (from 1 = *strongly disagree* to 5 =

Table 1. Reliability estimates separated per group

Scale/factor	Cuban sample		German sample	
	With situation descriptions	Without situation descriptions	With situation descriptions	Without situation descriptions
	ω [95% CI]	ω [95% CI]	ω [95% CI]	ω [95% CI]
SJT: personal initiative	.61 [.50, .73]	.67 [.56, .78]	.66 [.54, .78]	.74 [.66, .82]
Self-rating: personal initiative	.76 [.68, .83]	.82 [.75, .88]	.87 [.82, .91]	.84 [.78, .89]
Chance to perform (general)	.87 [.81, .92]	.89 [.85, .92]	.72 [.61, .83]	.79 [.71, .87]
Chance to perform (PI)	.87 [.82, .92]	.89 [.84, .93]	.85 [.80, .91]	.86 [.80, .91]
Test-taking motivation	.76 [.69, .84]	.56 [.30, .82]	.82 [.75, .90]	.81 [.74, .88]
Face validity	.68 [.53, .84]	.75 [.66, .83]	.70 [.60, .80]	.68 [.58, .78]
Perceived predictive validity	.73 [.65, .82]	.79 [.72, .86]	.83 [.78, .89]	.81 [.75, .87]
Perceived knowledge of results	.40 [.18, .61]	.71 [.60, .83]	.55 [.36, .74]	.62 [.49, .75]
Positive affect	.83 [.74, .91]	.82 [.75, .90]	.82 [.75, .89]	.78 [.70, .86]

Note. $n = 100$ (Cuban sample, with situation descriptions), $n = 92$ (Cuban sample, without situation descriptions), $n = 110$ (German sample, with situation descriptions), $n = 103$ (German sample, without situation descriptions).

strongly agree). Note that after completing the SJT, test takers were provided with an explanation that they had worked on an SJT focused on personal initiative for office jobs. Furthermore, if necessary, we slightly adapted the wording of some items to make sure that all items can be answered in a meaningful way and were related to personal initiative. Reliability estimates were acceptable to good, with ω values ranging between .56 and .89 (see Table 1). The only exception was that in the Cuban sample with the situation descriptions, the scale for perceived knowledge of results had a rather low reliability of McDonald's $\omega = .40$ (notably, reliability did not differ substantially between conditions, with overlapping 95% confidence intervals for McDonald's ω).

Results

We examined the SJT's measurement invariance, reliability, and construct-related validity. First, we grouped the SJT items into three parcels of items, based on the content domains that Bledow and Frese identified in their factor analysis, namely (1) personal initiative directed at improving organizational functioning, (2) personal initiative directed toward improving one's working conditions, and (3) personal initiative that required overcoming the resistance of supervisors and colleagues. The results from the 4-group CFA are presented in Table 2, and for analyses with all items as indicators, we refer to the electronic supplementary. As Mardia's test indicated no multivariate normality for all groups, we

used the robust MLR estimator for the CFA (see ES for details). We note that to avoid a just-identified model and biased estimates, SJT and self-rating on personal initiative were tested at the same time. Then, we applied the multigroup alignment procedure to evaluate measurement invariance multigroup factor analysis alignment (e.g., Marsh et al., 2018), as our goal was to make unbiased mean comparisons. Alignment is an alternative procedure to evaluate measurement invariance, without exact invariance tests as in traditional invariance testing. Similar to rotation algorithms in factor analysis, this approach optimizes a factor model to be sufficient for factor mean comparisons, i.e., it minimizes non-invariance between loadings and intercepts. Thereby, minor measurement differences across groups, which are common in cross-cultural research, are assumed and adjusted for, without fully rejecting a medium-fitting CFA model. In sum, the alignment approach produces estimates for the proportion of noninvariant parameters to inform decisions about the adequacy of mean comparisons similar to traditional invariance testing (e.g., Fischer & Karl, 2019; Luong & Flake, 2023).

Following the recommendations by Luong and Flake (2023), we started by testing a multigroup CFA model for configural invariance. In this model, we tested SJT parcels and self-ratings on personal initiative in a joint model to avoid a just-identified model and biased fit estimates. We achieved configural invariance, as suggested by a good fit of the 4-group model with $\chi^2(136) = 195.78, p = .001$; CFI = .944, RMSEA = .066 (90% CI [.045, .085]), SRMR = .062. In the second step, we evaluated for metric (loadings) and scalar (intercept) invariance of the SJT items via alignment

Table 2. Results from the CFA separated per group

Items	Cuban sample		German sample	
	With situation descriptions	Without situation descriptions	With situation descriptions	Without situation descriptions
	λ (SD)	λ (SD)	λ (SD)	λ (SD)
SJT: personal initiative				
Parcel 1	.29 (.50)	.39 (.51)	.44 (.59)	.43 (.53)
Parcel 2	.21 (.25)	.27 (.38)	.29 (.39)	.42 (.55)
Parcel 3	.64 (.78)	.79 (.95)	.60 (.70)	.78 (.85)
Self-rating: personal initiative				
Item 1	.48 (.61)	.45 (.72)	.74 (.80)	.66 (.74)
Item 2	.33 (.53)	.48 (.77)	.57 (.67)	.38 (.48)
Item 3	.58 (.75)	.68 (.82)	.75 (.81)	.73 (.76)
Item 4	.58 (.69)	.74 (.83)	.73 (.77)	.81 (.86)
Item 5	.31 (.41)	.35 (.42)	.62 (.69)	.65 (.69)
Item 6	.47 (.49)	.35 (.37)	.44 (.44)	.47 (.45)
Item 7	.27 (.37)	.32 (.47)	.55 (.61)	.47 (.51)

Note. $n = 100$ (Cuban sample, with situation descriptions), $n = 92$ (Cuban sample, without situation descriptions), $n = 110$ (German sample, with situation descriptions), $n = 103$ (German sample, without situation descriptions). Standardized λ in brackets. All loadings were significant, $p < .001$.

optimization. As suggested by Muthén and Asparouhov (2014), we evaluated the R^2 (in which 1 indicates complete invariance and 0 indicates complete noninvariance) and a threshold of $\leq 25\%$ noninvariant parameters as indicators for achieving invariance (see Luong & Flake, 2023). The results suggested metric invariance, with $R^2 = .98$ and 0% noninvariant loadings, and scalar invariance with $R^2 = .68$ and 25% noninvariant intercepts for the SJT items. Thus, correlations and means could be compared across the four groups.

Second, reliability estimates for the SJT were acceptable (with McDonald's ω values ranging between .61 and .74) and in line with previous meta-analytic findings (e.g., Kasten & Freund, 2016). As 95% confidence intervals of the McDonald's ω estimates overlapped, reliability did not significantly differ between the four groups (see Table 1).

Finally, supporting construct-related validity, the (latent) correlation between the SJT and the self-rating of personal initiative was strong in all groups, $\varphi_{\text{Cuban sample with situation descriptions}} = .55, p < .001$, $\varphi_{\text{Cuban sample without situation descriptions}} = .50, p < .001$; $\varphi_{\text{German sample with situation descriptions}} = .71, p < .001$, and $\varphi_{\text{German sample without situation descriptions}} = .61, p < .001$. To test for differences in construct-related validity, we used robust Monte Carlo confidence intervals: They indicated no significant differences between the German and Cuban samples (with situation description: $\Delta\varphi = -.16, 95\% \text{ CI } [-0.48, 0.15]$; without situation description: $\Delta\varphi = -.12, 95\% \text{ CI } [-0.40, 0.17]$) or due to situation availability (Cuban: $\Delta\varphi = .06, 95\% \text{ CI } [-0.28, 0.38]$; German: $\Delta\varphi = .10, 95\% \text{ CI } [-0.17, 0.37]$). In sum, we found similar psychometric properties in both groups.

In H1, we proposed that face validity, perceived predictive validity, opportunity to perform, perceived

knowledge of results, positive affect, and test-taking motivation were higher for German participants. To test our hypothesis, we conducted a two-way MANOVA [factors: culture (I): Cuba versus Germany; situation description (II): with versus without]. The omnibus test revealed a significant difference between the cultures, Wilk's $\lambda = .82, F(7, 395) = 12.33, p < .001$. However, neither the availability of a situation description, Wilk's $\lambda = .99, F(7, 395) = 0.62, p = .739$, nor their interaction revealed significant group differences, Wilk's $\lambda = .99, F(7, 395) = 0.80, p = .590$. Specifically, Cuban participants rated the face validity of the SJT higher than the German sample, $\eta^2 = .08, F(1, 401) = 32.76, p < .001$, attributed more predictive validity to the SJT, $\eta^2 = .02, F(1, 401) = 7.68, p = .006$, and indicated a higher perceived knowledge of results, $\eta^2 = .02, F(1, 401) = 7.21, p = .008$. The Cuban sample also showed higher test-taking motivation, $\eta^2 = .03, F(1, 401) = 11.06, p < .001$, which is why we controlled for test-taking motivation in the following analyses. We found no differences between Cuban and German participants regarding positive affect, $\eta^2 = .003, F(1, 401) = 1.01, p = .315$, and the perceived chance to perform, $\eta^2 = .002, F(1, 401) = 0.64, p = .423$ (for an overview, see Table 3 and Figure 1).

In H2, we suggested that situation descriptions are more important for SJT performance for Cuban participants than for German test takers. To test this, we conducted a two-way fixed-effects ANOVA with the factors culture (Cuba vs. Germany) and availability of situation description (with vs. without); we included test-taking motivation, weekly working hours, and job experience as control variables. We found that SJT scores were higher in the Cuban sample, $F(1, 398) = 17.74, p < .001$, and the availability of situation descriptions was positively related to SJT scores, $F(1, 398) = 33.57, p < .001$. Importantly, for H2, the interaction

Table 3. Two-way MANOVA for group differences between Cuba and Germany and the conditions with and without situation description

	Cuba (<i>n</i> = 198)		Germany (<i>n</i> = 213)		<i>F</i>	<i>p</i>	Situation (<i>n</i> = 210)		No situation (<i>n</i> = 195)		<i>F</i>	<i>p</i>	Hypotheses <i>df</i>	Error <i>df</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
Applicant perceptions														
<i>F.</i> validity	3.50	0.58	3.23	0.35	32.76	***	3.36	0.48	3.34	0.50	0.18	.670	1	401
Pos. affect	3.25	0.93	3.35	0.96	1.01	.315	3.37	0.97	3.23	0.91	2.17	.141	1	401
C. t. p. (General)	3.22	0.78	3.23	0.73	0.02	.883	3.29	0.73	3.16	0.77	3.10	.079	1	401
C. t. p. (PI)	3.21	0.69	3.27	0.80	0.64	.423	3.30	0.73	3.18	0.77	2.56	.111	1	401
Knowledge of results	3.20	0.55	3.03	0.70	7.21	.008**	3.13	0.60	3.09	0.68	0.35	.553	1	401
Predictive validity	3.10	0.60	2.90	0.79	7.68	.006**	3.03	0.70	2.96	0.72	0.82	.365	1	401
Test motivation	3.94	0.50	3.73	0.75	11.06	***	3.84	0.66	3.81	0.65	0.27	.602	1	401

Note. *M* and *SD* are used to represent mean and *SD*, respectively. C. t. p. represents participants' perceived chance to perform in the SJT. *F* statistics and degrees of freedom are reported.

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

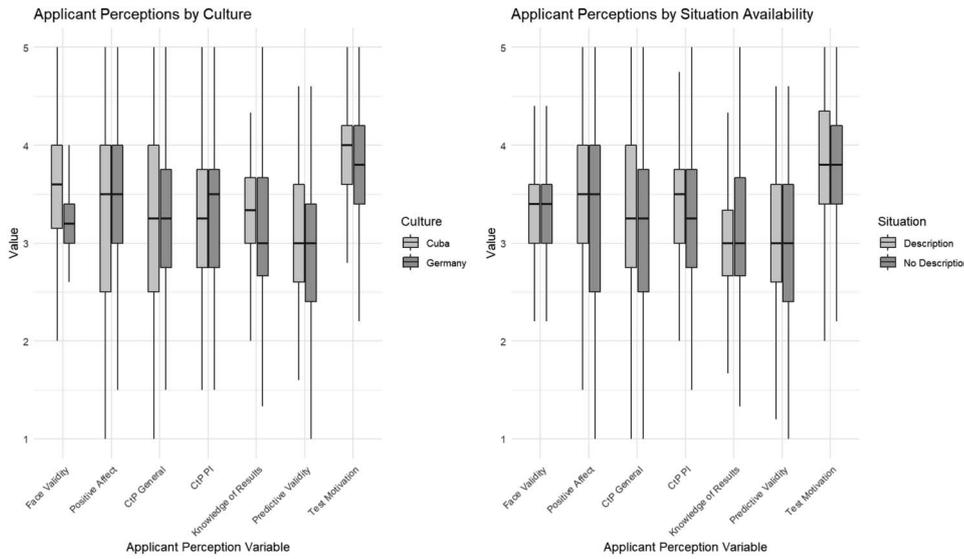


Figure 1. Comparison of applicant perception scales. The length of the boxplots depicts the middle 50% of the scores in the respective dependent variables between the conditions. CtP = chance to perform, PI = personal initiative.

Table 4. Fixed-effects ANOVA results using SJT test scores as the criterion

Predictor	Sum of squares	df	Mean square	F	p	η^2_{partial}	η^2_{partial} 90% CI [LL, UL]
Test motivation	0.44	1	0.44	5.21	.023	.01	
Job experience	0.85	1	0.85	9.99	.002**	.02	[.01, .05]
Working hours	0.02	1	0.02	0.28	.595	.00	[.00, .01]
Culture	1.50	1	1.50	17.74	***	.04	[.02, .08]
Situation	2.85	1	2.85	33.57	***	.08	[.04, .12]
Culture × Situation	0.01	1	0.01	0.12	.731	.00	[.00, .01]
Error	33.75	398	0.08				

Note. LL and UL represent the lower limit and upper limit of the partial η^2 confidence interval, respectively.
 ** $p < .01$. *** $p < .001$.

effect between the availability of situation descriptions and culture was not significant, $F(1, 398) = 0.12, p = .713$ (for details, see Table 4).

We also conducted analyses on the item level: We found that for five of 11 items (i.e., 45%) it did not make a difference whether the situation description was available, controlled for the effect of the culture (see Table 5). Furthermore, the culture did not have a significant effect on the item score for 64% (7 of 11) of the items, controlled for the effect of the situation. Notably, we also found no interaction effect between culture and situation availability on the item level (see Table 5). To test item-level mean differences per country, we conducted pairwise t tests (see Table 6), where α levels were Bonferroni-corrected ($p/\text{number of tests} = .05/11 = .0045$; Cabin & Mitchell, 2000). We found that excluding the situation description did not make a difference for 55% of the items (36% on the uncorrected α -level of .05) in the German sample and in the Cuban sample for 36% (36% on the uncorrected α -level).

Discussion

In the current study, we examined whether an SJT that was developed and validated in Germany could be transferred to the culture of Cuba. Thereby, we tested the generalizability of this SJT across cultural boundaries. Furthermore, we aimed to better understand the role of SJT item stems: We examined whether the finding that situation descriptions do not necessarily affect SJT item performance (e.g., Krumm et al., 2015; Schäpers, Mussel, et al., 2020) translates to other cultures. First, regarding psychometric properties, we found rather small differences between Cuba and Germany: Measurement invariance, construct-related validity, and reliability did not substantially differ between both cultures. Surprisingly though, test takers from Cuba did not perceive the SJT situations as more unconventional; they rated the face validity of the SJT higher than the German participants, attributed more predictive validity to the SJT, and indicated a higher perceived knowledge of

Table 5. Item-level main effects and interaction effect of culture and availability of situation description

SJT item number	Culture			Situation			Culture*Situation		
	<i>F</i>	<i>p</i>	η^2_{partial}	<i>F</i>	<i>p</i>	η^2_{partial}	<i>F</i>	<i>p</i>	η^2_{partial}
1	39.49	**	.10	36.40	**	.08	0.00	.964	.00
3	0.01	.931	.00	14.63	**	.04	0.42	.520	.00
4	8.49	.003*	.02	15.88	**	.04	0.94	.334	.00
5	16.81	**	.04	7.51	.006	.02	4.05	.045	.01
6	12.20	**	.03	20.52	**	.05	0.65	.421	.00
7	50.81	**	.11	5.99	.047	.01	4.89	.028	.01
8	0.25	.618	.00	3.07	.016	.01	3.07	.081	.01
9	6.40	.012	.01	115.02	**	.22	0.11	.740	.00
10	3.90	.049	.01	13.48	**	.03	2.32	.129	.01
11	1.11	.292	.00	6.28	.012	.02	0.17	.682	.00
12	0.08	.768	.00	0.36	.536	.00	3.41	.066	.01

Note. The results from two-way ANOVAs using SJT item scores as dependent variables. The results controlled for the effects of test motivation, job experience, and weekly working hours. Effects of control variables are not shown in this table.

* $p < .0045$. ** $p < .001$ (p level adjusted to account for α inflation: $p/\text{number of tests} = .05/11 = .0045$). Item 2 was removed due to a technical admission error in the Cuban sample.

Table 6. Item-level effects of the availability of situation descriptions per culture

SJT item number	Cuba				Germany			
	Cohen's <i>d</i>	<i>t</i>	<i>df</i>	<i>p</i>	Cohen's <i>d</i>	<i>t</i>	<i>df</i>	<i>p</i>
1	0.56	3.73	170.27	*	0.61	4.46	210.75	*
3	0.27	1.99	166.77	.05	0.46	3.31	207.31	*
4	0.45	3.34	189.94	*	0.31	2.27	210.38	.02
5	0.52	3.53	185.27	*	0.09	0.67	210.82	.50
6	0.42	3.08	188.49	*	0.42	3.06	210.64	*
7	-0.07	-0.57	190.00	.57	-0.44	-3.21	206.08	*
8	-0.51	-3.33	186.26	*	-0.05	-0.37	210.69	.71
9	1.05	7.24	189.47	*	1.10	7.91	178.58	*
10	0.58	4.09	184.17	*	0.22	1.62	206.58	.11
11	0.17	1.11	188.45	.27	0.30	2.19	205.24	.03
12	-0.14	-0.94	189.87	.35	0.23	1.71	211.00	.09

Note. One-sided *t* tests. Higher effect sizes reflect more correct answers on items with situation descriptions compared with items without situation descriptions.

* $p < .0045$ (p level adjusted to account for α inflation: $p/\text{number of tests} = .05/11 = .0045$). Item 2 was removed due to a technical admission error in the Cuban sample.

results. Second, regarding the impact of situation descriptions, we found no interaction between situation availability and culture; that is, the effect of situation availability on SJT performance did not substantially depend on the test takers' cultural background.

These findings have various implications for our understanding of SJT psychometrics. First, we found evidence that SJTs that were developed in Western cultures might be successfully transferred to other cultures, i.e., they can measure the intended construct validity. Since SJTs are classified as highly contextual measurements, applying them to another language and culture is a major challenge (Herde et al., 2019; Lievens, 2006). Thus, it is noteworthy that we report similar psychometric

properties for this SJT in Cuba and Germany. This extends previous findings that reported satisfactory but mixed evidence (e.g., Lievens et al., 2015). We suppose that this could depend on the SJT in question: We applied a construct-based SJT (Bledow & Frese, 2009), focusing on a single construct (i.e., personal initiative) instead of assessing a *broad* skill (e.g., teamwork competencies), as is the case with many established SJTs (see Lievens et al., 2021).

Second, we replicated the finding that situation descriptions do not necessarily affect SJT item performance or applicant perceptions, as we found that almost half of the items were not significantly easier to solve when a situation description was provided, although previous

studies showed even stronger indication of the context independency of SJTs with up to 70% of the items not changing in the results when the situation description was omitted (e.g., Krumm et al., 2015). This is notable given that Cuban test takers are not as familiar with typically German work-related situations. Thus, we provide further evidence that test takers do not always rely on the presented situation descriptions. One explanation might be that the response options already provide enough context information for test takers to infer the intended critical incident; for first evidence, see Freudenstein et al. (2020). On a different note, Lievens and Motowidlo (2016) argued that SJT performance represents a general domain knowledge instead of a situational judgment. Following this argument, context information is less relevant, and test takers use more general solution strategies when working on an SJT item. Third, we applied an experimental test validation strategy (for an overview, see Krumm et al., 2019): We showed that this strategy can also be applied to cross-cultural research questions and is a valuable extension of common correlative validation approaches.

Limitations and Conclusion

Our study was solely based on one SJT. Thus, we call for future cross-cultural research regarding SJT psychometrics (for further examples, see Herde et al., 2019). Furthermore, we only compared two cultures. To improve the generalizability of our findings, it is necessary to compare original and adapted versions between further cultures. Second, we did not recruit participants in a high-stakes assessment setting. Thus, one might argue that participants behave differently in real job-selection situations. Third, as we used parcels to test measurement invariance, investigations on the item level need to be seen with caution (for additional analyses, see the electronic supplementary). Nonetheless, we would like to mention that our approach followed the procedure of the test authors (see Bledow & Frese, 2009). Finally, our manipulation consisted of removing the entire situation from the SJT. We call for future research to use a more fine-grained approach by manipulating individual sections of the situation (e.g., location, acting persons, or the problem statement). In the same vein, we also call for more qualitative approaches (e.g., think-aloud technique) that evaluate the response evaluation processes in more detail.

In the end, we found initial evidence that a construct-driven SJT might be successfully transferred to other cultures. Furthermore, we contributed to the question about the role of situation description in SJTs by showing that SJT performance did not substantially depend on the test takers' cultural background.

References

- Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, 43(4), 695–716. <https://doi.org/10.1111/j.1744-6570.1990.tb00679.x>
- Banai, M., & Reisel, W. D. (2007). The influence of supportive leadership and job characteristics on work alienation: A six-country investigation. *Journal of World Business*, 42(4), 463–476. <https://doi.org/10.1016/j.jwb.2007.06.007>
- Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, 54(2), 387–419. <https://doi.org/10.1111/j.1744-6570.2001.tb00097.x>
- Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, 62(2), 229–258. <https://doi.org/10.1111/j.1744-6570.2009.01137.x>
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1(3), 185–216. <https://doi.org/10.1177/13591045700010030>
- 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(3), 246–248. <https://www.jstor.org/stable/20168454>
- Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures. In N. D. Christiansen, & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 439–456). Routledge. <https://doi.org/10.4324/9780203526910.ch19>
- Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63(1), 83–117. <https://doi.org/10.1111/j.1744-6570.2009.01163.x>
- Díaz Bravo, C., & Pañellas Alvarez, D. (2019). Cuba y Estados Unidos: Autoimagen e identidad nacional [Cuba and United States: Self image and national identity]. *Estudios del Desarrollo Social: Cuba y América Latina*, 7(2), 48–60.
- Fausey, C. M., Long, B. L., Inamori, A., & Boroditsky, L. (2010). Constructing agency: The role of language. *Frontiers in Psychology*, 1, Article 162. <https://doi.org/10.3389/fpsyg.2010.00162>
- Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, 10, Article 1507. <https://doi.org/10.3389/fpsyg.2019.01507>
- Frese, M., Fay, D., Hilburger, T., Leng, K., & Tag, A. (1997). The concept of personal initiative: Operationalization, reliability and validity in two German samples. *Journal of Occupational and Organizational Psychology*, 70(2), 139–161. <https://doi.org/10.1111/j.2044-8325.1997.tb00639.x>
- Freudenstein, J.-P., Schäpers, P., Römer, L., Mussel, P., & Krumm, S. (2020). Is it all in the eye of the beholder? The importance of situation construal for situational judgment test performance. *Personnel Psychology*, 73(4), 669–700. <https://doi.org/10.1111/peps.12385>
- Galati, D., Manzano, M., Roca, M., Sotgiu, I., & Fassio, O. (2004). Emotions and everyday life in Cuba. *Psychology and Developing Societies*, 16(2), 139–157. <https://doi.org/10.1177/097133360401600204>
- Geisinger, K. F. (2003). Testing and assessment in cross-cultural psychology. In J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (pp. 95–117). John Wiley & Sons.
- Herde, C., Lievens, F., Solberg, E. G., Strong, M. H., & Burkholder, G. J. (2019). Situational judgment tests as measures of 21st century skills: Evidence across Europe and Latin America. *Journal of Work and Organizational Psychology*, 35(2), 65–74. <https://doi.org/10.5093/jwop2019a8>

- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Sage.
- Jones, P. S., Lee, J. W., Phillips, L. R., Zhang, X. E., & Jaceldo, K. B. (2001). An adaptation of Brislin's translation model for cross-cultural research. *Nursing Research*, 50(5), 300–304. <https://doi.org/10.1097/00006199-200109000-00008>
- Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment*, 22(3), 168–176. <https://doi.org/10.1027/1015-5759.22.3.168>
- Kanning, U. P., & Kuhne, S. (2006). Social desirability in a multimodal personnel selection test battery. *European Journal of Work and Organizational Psychology*, 15(3), 241–261. <https://doi.org/10.1080/13594320600625872>
- Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of situational judgment tests (SJTs). *European Journal of Psychological Assessment*, 32(3), 230–240. <https://doi.org/10.1027/1015-5759/a000250>
- Krumm, S., Hüffmeier, J., & Lievens, F. (2019). Experimental test validation: Examining the path from test elements to test performance. *European Journal of Psychological Assessment*, 35(2), 225–232. <https://doi.org/10.1027/1015-5759/a000393>
- Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How “situational” is judgment in situational judgment tests? *Journal of Applied Psychology*, 100(2), 399–417. <https://doi.org/10.1037/a0037674>
- Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 279–300). Lawrence Erlbaum.
- Lievens, F., Corstjens, J., Sorrel, M. Á., Abad, F. J., Olea, J., & Ponsoda, V. (2015). The cross-cultural transportability of situational judgment tests: How does a US-based integrity situational judgment test fare in Spain?. *International Journal of Selection and Assessment*, 23(4), 361–372. <https://doi.org/10.1111/ijsa.12120>
- Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, 9(1), 3–22. <https://doi.org/10.1017/iop.2015.71>
- Lievens, F., Schäpers, P., & Herde, C. N. (2021). Situational judgment tests: From low-fidelity simulations to alternative measures of personality and the person-situation interplay. In D. Wood, S. J. Read, P. D. Harms, & A. Slaughter (Eds.), *Measuring and modeling persons and situations* (pp. 285–311). Academic Press. <https://doi.org/10.1016/B978-0-12-819200-9.00017-X>
- Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*, 28(4), 905–924. <https://doi.org/10.1037/met0000441>
- Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*, 23(3), 524–545. <https://doi.org/10.1037/met0000113>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. <https://doi.org/10.1037/a0028085>
- Mussel, P., Gatzka, T., & Hewig, J. (2018). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment*, 34(5), 328–335. <https://doi.org/10.1027/1015-5759/a000346>
- Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, 5, Article 978. <https://doi.org/10.3389/fpsyg.2014.00978>
- Nohr, L., Lorenzo Ruiz, A., Sandoval Ferrer, J. E., & Buhlmann, U. (2021). Mental health stigma and professional help-seeking attitudes a comparison between Cuba and Germany. *PLoS ONE*, 16(2), Article e0246501. <https://doi.org/10.1371/journal.pone.0246501>
- Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teacher*, 38(1), 3–17. <https://doi.org/10.3109/0142159x.2015.1072619>
- Reali, F., Spivey, M., Tyler, M., & Terranova, J. (2006). Inefficient conjunction search made efficient by concurrent spoken delivery of target identity. *Perception & Psychophysics*, 68(6), 959–974. <https://doi.org/10.3758/bf03193358>
- Schäpers, P., Freudenstein, J.-P., Mussel, P., Lievens, F., & Krumm, S. (2020). Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests. *Journal of Research in Personality*, 87(8), 357–375. <https://doi.org/10.1016/j.jrp.2020.103963>
- Schäpers, P., Lievens, F., Freudenstein, J.-P., Hüffmeier, J., König, C. J., & Krumm, S. (2020). Removing situation descriptions from situational judgment test items: Does the impact differ for video-based versus text-based formats?. *Journal of Occupational and Organizational Psychology*, 93(2), 472–494. <https://doi.org/10.1111/joop.12297>
- Schäpers, P., Mussel, P., Lievens, F., König, C. J., Freudenstein, J.-P., & Krumm, S. (2020). The role of situations in situational judgment tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology*, 105(8), 800–818. <https://doi.org/10.1037/ap10000457>
- Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, 46(1), 49–76. <https://doi.org/10.1111/j.1744-6570.1993.tb00867.x>
- Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, 54(2), 119–135. <https://doi.org/10.1016/j.erap.2003.12.004>

History

Received December 6, 2022

Revision received November 9, 2023

Accepted November 10, 2023

Published online January 3, 2024

Section: I/O Psychology

Acknowledgments

Philipp Schäpers, Franz W. Mönke, and Henrik Heinemann thank the State of North Rhine-Westphalia's Ministry of Economic Affairs, Industry, Climate Action, and Energy as well as the Exzellenz Start-up Center.NRW program at the REACH – EUREGIO Start-Up Center for their kind support of our work. The authors thank Celeste Brennecke for proofreading the manuscript.

Publication Ethics

The scientific committee at the Faculty of Psychology of the Universidad de La Habana approved our procedure.

Authorship

Philipp Schäpers: conceptualization, methodology, formal analysis, writing – original draft, supervision; Henrik Heinemann: methodology, investigation, formal analysis, writing – original draft; Daybel Pañellas Alvarez: investigation; Laura Nohr: investigation;

Franz W. Mönke: methodology, investigation, formal analysis, writing – original draft.

Open Science

Open Data: The authors confirm that there is sufficient information for an independent researcher to reproduce all the reported results. All data are available at <https://bit.ly/3gQRHlh>.

Preregistration and Analysis Plan: Hypotheses, study design, and analyses of this study are available at https://aspredicted.org/9JX_8PL.

Data and analysis code are available in the electronic supplementary (ES) at the Open Science Framework (<https://bit.ly/3gQRHlh>).

Funding

This cooperation was supported by ERASMUS staff mobility. Open access publication was enabled by the University of Münster, Germany.

ORCID

Philipp Schäpers

 <https://orcid.org/0000-0002-8270-5105>

Henrik Heinemann

 <https://orcid.org/0000-0003-1651-9045>

Laura Nohr

 <https://orcid.org/0000-0002-3798-0909>

Franz W. Mönke

 <https://orcid.org/0000-0002-6634-0193>

Philipp Schäpers

Department of Psychology

University of Münster

Fliednerstraße 21

48149 Münster

Germany

philipp.schaepers@uni-muenster.de