Journal of Work and Organizational Psychology (2019) 35(2) 65-74

Journal of Work and Organizational Psychology

https://journals.copmadrid.org/jwop

Situational Judgment Tests as Measures of 21st Century Skills: Evidence across Europe and Latin America

Christoph N. Herde[a], Filip Lievens[b], Emily G. Solberg[c], Jan L. Harbaugh[c], Mark H. Strong[d], and Gary J. Burkholder[e]

[a]Ghent University, Belgium; [b]Singapore Management University, Singapore; [c]SHL, USA; [d]Strong Talent Strategies, Houston, TX, USA; [e]Walden University, Minneapolis, MN, USA

ARTICLE INFO

Article history:
Received 30 December 2018
Accepted 25 March 2019
Available online 21 June 2019

Keywords:
Situational judgment test
21st century skills
Measurement invariance

ABSTRACT

Over the years, various governmental, employment, and academic organizations have identified a list of skills to successfully master the challenges of the 21st century. So far, an adequate assessment of these skills across countries has remained challenging. Limitations inherent in the use of self-reports (e.g., lack of self-insight, socially desirable responding, response style bias, reference group bias, etc.) have spurred on the search for methods that could complement or even substitute self-report inventories. Situational judgment tests (SJTs) have been proposed as one of the complements/alternatives to the traditional self-report inventories. SJTs are low-fidelity simulations that confront participants with multiple domain-relevant situations and request to choose from a set of predefined responses. Our objectives are twofold: (a) outlining how a combined emic-etic approach can be used for developing SJT items that can be used across geographical regions and (b) investigating whether SJT scores can be compared across regions. Our data come from Laureate International Universities (N = 5,790) and comprise test-takers from Europe and Latin America who completed five different SJTs that were developed in line with a combined emic-etic approach. Results showed evidence for metric measurement invariance across participants from Europe and Latin America for all five SJTs. Implications for the use of SJTs as measures of 21st century skills are discussed.

Los tests de juicio situacional como medida de las habilidades para el siglo XXI: evidencia en toda Europa y América Latina

Palabras clave:
Test de juicio situacional
Habilidades para el siglo XXI
Invarianza de la medición

RESUMEN

A lo largo de los años, varias organizaciones gubernamentales de empleo y académicas han identificado una lista de habilidades para superar con éxito los desafíos del siglo XXI. Hasta ahora, una evaluación adecuada de estas habilidades en los países ha continuado siendo un reto. Las limitaciones inherentes al uso de autoinformes (p. ej., falta de autoconocimiento, respuestas socialmente deseables, sesgo en el estilo de respuesta, sesgo del grupo de referencia, etc.) han estimulado la búsqueda de métodos que puedan complementar o incluso sustituir inventarios de autoinforme. Los tests de juicio situacional (TJS) se han propuesto como uno de los complementos/alternativas a los inventarios tradicionales de autoinforme. Los TJS son simulaciones de baja fidelidad que enfrentan a los participantes con múltiples situaciones de dominio relevantes y solicitan elegir entre un conjunto de respuestas predefinidas. Tenemos un doble objetivo: (a) explicar cómo se puede utilizar un enfoque *emic-etic* combinado para desarrollar ítems de TJS que se puedan emplear en todas las regiones geográficas y (b) investigar si las puntuaciones de los TJS se pueden comparar entre regiones. Nuestros datos provienen de las *Laureate International Universities* (N = 5,790) y están compuestos por examinandos de Europa y América Latina que cumplimentaron cinco TJS diferentes que se desarrollaron de acuerdo a un enfoque *emic-ethic*. Los resultados mostraron la existencia de invarianza en la medición en los participantes de Europa y América Latina para los cinco TJS. Se discuten las implicaciones para el uso de TJS como medida para detectar habilidades en el siglo XXI.

Cite this article as: Herde, C. N., Lievens, F., Solberg, E. G., Harbaugh, J. L., 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*, 65-74. https://doi.org/10.5093/jwop2019a8 [Antonio García-Izquierdo and David Aguado were the guest editors for this article].

Correspondence: christoph.herde@ugent.be (C. N. Herde).

ISSN:1576-5962/© 2019 Colegio Oficial de Psicólogos de Madrid. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Situational Judgment Tests as Measures of 21st Century Skills: Evidence across Europe and Latin America

Since several decades, various educational and (non)profit organizations around the globe have compiled lists of skills needed for the next generation to survive in an ever changing, turbulent, and complex world. Although the final lists of these large-scale international efforts often differ in their name ("survival skills", "21st century skills") and content, they all share the characteristic that the skills identified go beyond technical and functional aptitude. The most common examples of such 21st century skills are, therefore, collaboration and teamwork, creativity and imagination, critical thinking, and problem solving (see, for overviews, Binkley et al., 2012; Geisinger, 2016).

Besides identifying the list of 21st century skills, an equally important issue deals with how these skills are best measured. Specifically, challenges deal with using a methodology that does not lead to biases and that enables comparing the results obtained across the various geographical regions. Along these lines, it is of pivotal importance that measurement effects do not cloud the standing of the regions on the 21st century skills (constructs). In the past, self-reports were typically used for determining people's standing on each of the skills. However, the self-report methodology suffers from various pitfalls. One drawback is that self-reports assume people possess the necessary self-insight to rate themselves on each of the statements that operationalize the 21st century skills. Another drawback is that people tend to engage in response distortion in that they might overstate how they score on the statements (socially desirable responding). Other documented limitations relate to response style bias (extreme responding that differs across groups, such as different cultures; e.g., Hui & Triandis, 1989; Johnson, Kulesa, Cho, & Shavitt, 2005) or reference group bias (responding that is dependent on the chosen group of reference, such as one's own cultural group; e.g., Heine, Lehman, Peng, & Greenholtz, 2002).

These limitations have resulted in the search for other methods for measuring 21st century skills (Kyllonen, 2012; see also Ainley, Fraillon, Schulz, & Gebhardt, 2016; Care, Scoular, & Griffin, 2016; Ercikan & Oliveri, 2016; Greiff & Kyllonen, 2016; Herde, Wüstenberg, & Greiff, 2016; Lucas, 2016). In PISA (OECD, 2014), three such approaches were suggested (for a summary, see Kyllonen, 2012). The first method dealt with the use of anchoring vignette items. Anchoring vignette items first ask respondents to evaluate several other targets on a specific target construct. Only afterwards, a respondent provides a self-rating on the target construct. The respondent's self-rating is then rescaled based upon the evaluation standards that are extracted from the other ratings (e.g., Hopkins & King, 2010). As a second approach, forced choice methods were proposed. Forced-choice items do not ask respondents to evaluate isolated statements about themselves on a Likert-scale. Instead, they confront respondents with a choice between options that are intended to be of similar social desirability. Recent research attested to the broad applicability of forced choice items (Brown & Maydeu-Olivares, 2011; Stark, Chernyshenko, & Drasgow, 2004). Third, situational judgment tests (SJTs) were proposed. SJTs confront respondents with multiple, domain-relevant situations and request to choose from a set of predefined responses (Motowidlo, Dunnette, & Carter, 1990).

Importantly, these approaches aim to alleviate the limitations inherent in the typical self-report inventories, while at the same time ensuring that the average ratings on the 21st century skills can be compared across geographical regions. Note that SJTs do not actually measure 21st century skills. Instead, SJTs assess people's procedural knowledge ("knowing what to do and how to do it") of engaging in behavior that operationalizes a given 21st century skill (Lievens, 2017; Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo, Hooper, & Jackson, 2006).

In this study, we focus on the use of SJTs as measures of 21st century skills. Our objectives are twofold. First, we outline how a combined emic-etic approach can be used for developing SJT items that can be used across geographical regions. Second, we investigate whether SJT scores derived from a SJT that was developed in line with a combined emic-etic approach can indeed be compared across regions. We do so by conducting analyses of measurement invariance across regions of Europe and Latin America. Analyses of measurement invariance reveal whether different (regional or cultural) groups interpret test items in the same way and attribute the same meaning to them. Therefore, analyses of measurement invariance are crucial to disentangle measurement effects from true score differences between (regional or cultural) groups (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000).

Our study is situated in an educational context. We use the data from Laureate International Universities, which is a global network of universities that, at the time of the study, operated in 25 countries and had over one million students globally. Similar to the efforts described above, Laureate International Universities started in 2015 to identify, define, and measure foundational competencies and behavioral skills required by graduating students to be successful in entry-level professional jobs across industries and geographical regions. SJT items were also developed to assess those foundational competencies. On the basis of the SJT scores, students receive feedback regarding their strengths and weaknesses as well as actionable tips to help them improve. It is also important that regions can be compared on their average standing on the various competencies.

The structure of this paper is as follows: First, we shortly define SJTs and illustrate their most important characteristics. Second, we explain why these special characteristics of SJTs may pose problems for measurements across geographical regions. Third, we describe how a combined emic-etic approach of test development might serve to limit these problems. Fourth, we provide an empirical test of the combined emic-etic approach to develop SJTs to measure 21st century skills across geographical regions of Europe and Latin America. Fifth, we discuss our results and implications for further research and practice.

Study Background

SJTs: Definition, Characteristics, and Brief History

In SJTs, candidates are presented with short domain-relevant situational descriptions and various response options to deal with the situations. Upon reading the short situational descriptions, candidates are asked to pick one response option from a list, rank the response options ("What would you prefer doing most, least?"), or rate the effectiveness of these options (Motowidlo et al., 1990). Most SJTs still take the form of a written test because the scenarios are presented in a written format. In video-based or multimedia SJTs, a number of video scenarios describing a person handling a critical situation is developed (McHenry & Schmitt, 1994). Recently, organizations are also exploring 2D-animated, 3D-animated, and even avatar-based SJTs (see, for an overview, Weekley, Hawkes, Guenole, & Ployhart, 2015).

SJTs are not new inventions. Early SJT versions go back to before WWII. In 1990, Motowidlo and colleagues reinvigorated interest in SJTs. Since then, SJTs have become attractive selection instruments for practitioners who are looking for cost-effective instruments. As compared to other sample-based predictors, SJTs might be easily deployed via the internet in a global context due to their efficient administration (Ployhart, Weekley, Holtz, & Kemp, 2003). Moreover, in domestic employment contexts, SJTs have demonstrated adequate criterion-related and incremental validity and potential to reduce adverse impact (Christian, Edwards, & Bradley, 2010; McDaniel, Hartman, Whetzel, & Grubb III, 2007; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001).

SJTs in an International Context: Potential Problems

Although SJTs have been advanced as alternative method for assessing 21st century skills across geographical regions, such an outcome is far from assured. For example, Ployhart and Weekley (2006) mentioned the following key challenge: "it is incumbent on researchers to identify the cross-cultural generalizability – and limits – of SJTs… One might ask whether it is possible to create a SJT that generalizes across cultures. Given the highly contextual nature of SJTs, that poses a very interesting question." (p. 349). Indeed, SJT items are directly developed or sampled from the criterion behaviors that the test is designed to predict (Chan & Schmitt, 2002). Therefore, SJT items are highly contextualized because the situations are embedded in a particular context or situation that is representative of future tasks.

Lievens (2006) reviewed prior research on SJTs in a cross-cultural context and also identified SJT item characteristics that might affect the cross-cultural use of SJTs (see also Lievens et al., 2015). The contextualized nature of SJT items makes them particularly prone to cultural differences because the culture wherein one lives acts like a lens, guiding the interpretation of events and defining appropriate behaviors (Heine & Buchtel, 2009; Lytle, Brett, Barsness, Tinsley, & Janssens, 1995). This contextualized nature of SJTs might create boundary conditions for the use across geographical regions in at least four ways (Lievens, 2006). First, the contextualization in SJTs is shown in the kind of problem situations (i.e., the item stems) that are presented to candidates. When SJTs are used in an international context, the issue then becomes whether there are differences in terms of the situations (critical incidents) generated across regions. Some situations will simply not be relevant in one region, whereas they might be very relevant in another region (e.g., differences in organizing meetings across countries). Second, similar differences might occur on the level of how to react to the problem situation. That is, some response options might be relevant in one region, whereas they might not occur in another region. The meeting example can again be used here, with openly not agreeing with the boss being an unrealistic response option in some regions. Third, the effectiveness (scoring) of response options might vary across regions. Along these lines, Nishii, Ployhart, Sacco, Wiechmann, and Rogg, (2001) stated: "if a scoring key for a SJT is developed in one country and is based on certain cultural assumptions of appropriate or desirable behavior, then people from countries with different cultural assumptions may score lower on these tests. Yet these lower scores would not be indicative of what is considered appropriate or desirable response behavior in those countries". Fourth, the item-construct linkages might differ across regions. That is, a specific response option might be an indicator of a given construct in one region but an indicator of another construct in another region. For example, to decline a task assignment from a supervisor because of time constraints during a department meeting might indicate assertiveness or self-regulation in a culture low in power distance but might indicate impoliteness or rudeness in a culture high in power distance.

In short, these potential differences in the situations, response options, response option effectiveness, and item-construct linkages across geographical regions highlight that care should be taken to develop SJTs for measuring 21st century skills across regions. That is, strategies should be deployed for designing SJTs that alleviate these potential problems.

Strategies for SJT Design in a Cross-cultural Context: Emic, Etic, and Combined Emic-Etic Approach

In the search of strategies for dealing with potential threats to the cross-cultural transportability of SJTs, it is possible to borrow valuable insights from the large body of research in cross-cultural psychology. Generally, three possible approaches can be adopted for developing global (selection) instruments, namely an emic, an imposed etic, and a combined emic-etic approach (Berry, 1969, 1990; Headland, Pike, & Harris, 1990; Leong, Leung, & Cheung, 2010; Morris, Leung, Ames, & Lickel, 1999; Pike, 1967; Yang, 2000).

An indigenous or *emic approach* posits that tests should be developed and validated with the own culture as a point-of-reference. In the context of SJTs, an example is the study of Chan and Schmitt (2002). They developed an SJT for civil service positions in Singapore. This implied that the job analysis, the collection of situations, the derivation of response alternatives, the development of the scoring key, and the validation took place in Singapore. Chan and Schmitt (2002) found that in Singapore the SJT was a valid predictor for overall performance and had incremental validity over cognitive ability, personality, and job experience. This corresponds to the meta-analytic validity research base in the United States (Christian et al., 2010; McDaniel et al., 2007; McDaniel et al., 2001).

In this example, the development of the SJT ensured that the job relevant scenarios were derived from input of local subject matter experts. However, there are also drawbacks in the emic approach. As an indigenous approach implicates the use of different instruments for different countries, it is a costly and time-consuming strategy. In addition, a challenge for the country-specific emic approach is to contribute to the cumulative knowledge in a specific domain that typically centers around generalizable concepts (Leong et al., 2010; Morris et al., 1999).

Contrary to the emic approach, the *imposed etic approach* assumes that the same instrument can be applied universally across different cultures (Berry, 1969; Church & Lonner, 1998). So, according to the imposed etic approach, a selection procedure developed in a given country can be exported for use in other countries when guidelines for test translation and adaptation are taken into consideration (International Test Commission, 2001). Hence, the imposed etic approach represents an efficient strategy for cross-cultural assessment. However, the imposed etic approach is also not without limitations. Even when tests are appropriately translated and adapted, the test content of the transported instruments might reflect predominantly the culture from which the instrument is derived, thereby potentially omitting important emic aspects of the local culture (Cheung et al., 1996; Leong et al., 2010).

In light of these drawbacks, the effectiveness of the imposed etic approach for constructing international SJTs seems doubtful given the highly contextualized nature of SJT items. One study confirmed the problems inherent in using an imposed etic approach in contextualized instruments such as SJTs. Such and Schmidt (2004) validated an SJT in three countries. Results in a cross-validation sample showed that the SJT was valid in half of the countries, namely the United Kingdom and Australia. Conversely, it was not predictive in Mexico. These results suggest that effective behavior on the SJT was mainly determined in terms of what is considered effective behavior in two countries with a similar heritage (the United Kingdom and Australia).

Another study on the cross-cultural transportability of SJTs showed that an integrity SJT developed in the US was generally applicable to a Spanish population as well (Lievens, Corstjens et al., 2015). Most of the scenarios from the American SJT were rated to be realistic in the Spanish population, patterns of endorsements of various response options were mainly similar across cultures, the American scoring scheme correlated highly with Spanish scoring schemes and item-construct linkages also appeared to be comparable, because correlations between self-reports and SJT scores were found to be similar across cultures. In sum, evidence for the imposed etic approach for constructing international SJTs is mixed.

Yet, the emic-etic distinction should not be seen as a dichotomy. Rather, it constitutes a continuum (Church, 2001; Morris et al., 1999; Sahoo, 1993). Therefore, it is possible to combine these

cultural-general and cultural-specific approaches of international test development (Leong et al., 2010; Schmit, Kihm, & Robie, 2000), resulting in the *combined emic-etic approach*. In such a combined emic and etic approach, the instrument is developed with cross-cultural input. In the personality domain, we are aware of two prior projects that successfully applied the combined emic-etic approach. First, in the development of the Chinese Personality Assessment Inventory (CPAI; Cheung, Cheung et al., 2008; Cheung, Fan, Cheung, & Leung, 2008; Cheung et al., 1996) descriptions of personality were extracted from multiple sources (e.g., proverbs, everyday life, etc.) to identify personality constructs relevant to the Chinese culture. These local expressions were then compared to translations of imported measures of similar constructs. Large-scale tests of the inventory in China showed that there was substantial overlap between the CPAI and the Big Five, although there were also unique features (i.e., the interpersonal relatedness factor). As a second illustration, Schmit et al. (2000) developed a global personality inventory. Hereby the behavioral indicators (items) of personality constructs that were written by worldwide panels of local experts varied, while the broader underlying constructs were similar across countries. Construct-related validity studies provided support for the same underlying structure of the global personality inventory across countries.

So, as a result of a combined emic-etic approach, both universal and indigenous constructs are incorporated: the inclusion of culture-specific concepts produces within-culture relevance, while the measurement of universal concepts allows cross-cultural comparisons (Cheung et al., 1996). The combined emic-etic approach also enables to expand the interpretation of indigenous constructs in a broader cultural context.

In sum, prior studies have developed and used SJTs in various geographical regions. However, many applications were within-country examinations that attest to an indigenous (culture-specific/emic) approach. One study (Such & Schmidt, 2004) applied an imposed etic approach with the SJT not being valid in some countries. To avoid these problems, the combined emic-etic approach might serve as a potentially viable strategy for constructing sample-based selection procedures such as SJTs for use in cross-cultural applications. So far, no empirical studies have used or tested this combined emic-etic approach in sample-based selection procedures such as SJTs. This study starts to fill this key research and practice gap by using a combined emic-etic approach for constructing an SJT for assessing 21st century skills across geographical regions.

Method

Development and Validation of a Global Competency Framework

Laureate International Universities developed and validated a comprehensive framework of competencies that are required by graduating students to be successful in the workplace across geographical regions, industries, and jobs. In line with the combined emic-etic approach, cross-regional input was gained across all developmental steps to ensure that the competency framework was relevant across regions and cultures.

The development of the competency framework was based on various sources of information. These included best practices in competency modeling (Campion et al., 2011; Kurz & Bartram, 2002), content of competency frameworks from academic institutions and professional companies (e.g., Getha-Taylor, Hummert, Nalbandian, & Silvia, 2013; Lee, 2009; Lunev, Petrova, & Zaripova, 2013), internal research conducted by several institutions in the Laureate network, and data from various research partners. A draft competency framework was developed by integrating information from these sources and utilizing competency names and definitions from the SHL Universal Competency Framework (Bartram, 2012).

To ensure that the draft competency framework comprehensively covered competencies that were applicable and important across geographical regions, industries, and jobs, it was reviewed, refined, and approved by various groups. These groups included a global advisory council, consisting of eighteen members from regions represented in Laureate, two subject matter experts on competency modeling, and eighteen global focus groups that represented all regions, stakeholders (students/alumni, faculty/staff, academic leaders, and employers), and experts across disciplines. In total, the global focus group comprised of 86 participants.

Finally, two survey studies were conducted among Laureate's stakeholders across the network to evaluate and refine the competency framework. In Survey 1, 25,202 representatives across different stakeholders, roles, disciplines, and regions confirmed the importance of the competencies for entry-level professionals. In Survey 2, 10,420 of these representatives further reviewed and confirmed the individualist behaviors defined within each competency. The final competency framework includes 20 competencies. Further details about the competency framework, its development, and the global validation study are reported elsewhere (Strong, Burkholder, Solberg, Stellmack, & Presson, manuscript submitted for publication).

In this study, we focus on five core competencies that were identified in the global validation study as most important and critical for successful job performance of new professionals across geographical regions, industries, and jobs. These core competencies are achieving objectives, adapting to change, analyzing and solving problems, learning and self-development, and working well with others. The definitions of these competencies are provided in the Appendix.

SJT Item Design and Scoring

Analogous to the development of the competency framework, a combined emic-etic approach was applied to develop written SJT items with close-ended response format for the competencies. The development of the SJTs followed recommendations from Weekley, Ployhart, and Holtz (2006). We started with using the critical incident technique (Flanagan, 1954) to gain input for item development from subject matter experts. Given that the SJTs should assess competencies required of graduating students to be successful in the workplace, students, faculty/staff, administrators, alumni, and advisory committee members of Laureate institutions as well as employers served as subject matter experts. Representatives from these groups were invited to fill in an online survey to describe specific situations for a chosen competency, in which one student performed exceptionally well and another student performed exceptionally poorly. In total, 1,749 critical incidents were gathered from 564 respondents.

Three experienced test construction consultants drafted initial items. They compiled, reviewed and synthesized the critical incidents. Per competency, critical incidents and related examples for excellent and poor performance were converted into item stems and response options. Per item stem, five response options were generated that aimed to measure different levels of proficiency for the same competency.

Item stems and response options were written in a way to be applicable across different regions, industries, and jobs. To verify this, two global focus group panels reviewed all items and determined the scoring key. The panels consisted of 21 and 22 participants, respectively. Both panels represented similar numbers of representatives from all geographical regions, functional roles (Laureate faculty/staff and employers), and employers from different industries. Panelists reviewed items with special focus on realism and

face validity of depicted situations and response options within their geographical region and field of work. Potential issues were discussed and items were adapted, if necessary.

To set the scoring key per SJT, these panelists rated the effectiveness of each response option per item stem on a five-point scale (1 = *very ineffective*, 5 = *very effective*). In line with the consensus weighting method (see Chan & Schmitt, 1997), the average ratings were used to assign each response option a score of 1 through 5 points.

The items and related response options and scoring keys were further reviewed by assessment experts and employers. In total, twelve assessment experts (two per geographical region) with advanced degrees in Industrial/Organizational Psychology or a closely related discipline reviewed all items. Assessment experts provided feedback regarding item clarity or content from their own cultural perspective. Based upon this feedback, some items were slightly modified. Assessment experts also indicated whether each item appeared to tap into the respective competency. If at least half of the assessment experts indicated that an item did not appear to capture the targeted competency, the respective item was dropped. A final panel of fourteen employers reviewed all items. Again, this panel was formed by representatives from all global regions as well as from different industries and jobs.

After final minor item modifications, each of the competency specific SJTs constructed consisted of 21 items on average. Items had a behavioral tendency response instruction ("What would you do?"). For each item stem/scenario, participants were instructed to choose a response option they would most likely do and another response option they would least likely do. Participants could receive between 1 and 5 points for each choice. Therefore, scores could vary between 2 and 10 points per scenario.

All SJT items were translated from English into six additional languages. These additional languages were Latin American Spanish, European Spanish, Brazilian Portuguese, European Portuguese, French, and German. The rigorous translation process followed guidelines for translating tests (e.g., Van de Vijver, 2003), including repeated front and back translations by different translators.

Procedure and Sample

Laureate institutions invited their students to take part in this study to receive developmental feedback about their competency levels. The different SJTs were distributed across four different bundles that contained different competency specific SJTs. Students were invited to complete one bundle but could complete additional bundles to receive developmental feedback about further competencies. Within each bundle, students completed a random set of eight scenarios per competency specific SJT. Finally, students responded to demographic questions.

To assure that only valid data were analyzed, we removed data for several reasons. In a limited number of 24 cases, students started the same bundle twice. To exclude biases due to retest effects regarding the same competencies or scenarios, we excluded responses from the second bundle completion. For the same reason, we removed responses of eight students from the second access to any SJT of the same competency. Given that we were interested in cross-regional comparisons, we took care that participants understood the test items well. Hence, we removed data for 87 students that indicated to be "not comfortable" with the language in which they completed the SJTs. Further, we removed students' responses per scenario if they were made in less than twelve seconds (internal test runs had shown it was impossible to choose both a best and worst response per scenario in less than twelve seconds). Remaining sample sizes for our five core competencies did not justify analyses for the geographical regions of Africa, Asia, Oceania, or the US. Therefore, we focused our analyses on students from Europe and Latin America.

After data cleaning, a total of 5,790 students (53% female) from twenty different institutions provided valid responses to the competency specific SJTs (mean age = 22.63, *SD* = 5.09); 64% of the students resided in Europe, 36% in Latin America. In total, students came from eighteen different countries. The majority of European students resided in Turkey (30%), Portugal (20%), or Spain (17%). The majority of Latin American students lived in Mexico (34%), Chile (22%), or Brazil (18%). Each student chose to complete the SJTs in one of seven available languages. The majority of students completed the SJTs in English (32%), Latin American Spanish (29%), or European Portuguese (13%); 74 % of all students completed the SJTs in their dominant language; 72% of all students reported to be "very comfortable" with the language in which they completed the SJTs[1]. Students completed the SJTs either during their first (52%) or last year of study (48%) at the institution; 45% completed the SJTs in a proctored setting; 58% of students reported to have already gained some professional experience; 41% already completed

Table 1. Internal Consistencies, Means and Standard Deviations per Geographical Region by SJTs

	n	α	*M*	*SD*
		Achieving objectives (19 items)		
Europe and Latin America	3,666	.78	7.56	1.27
Europe	2,666	.79	7.57	1.23
Latin America	1,000	.78	7.53	1.38
		Adapting to change (20 items)		
Europe and Latin America	4,511	.69	7.58	1.17
Europe	3,586	.69	7.61	1.14
Latin America	925	.69	7.48	1.26
		Analyzing & solving problems (19 items)		
Europe and Latin America	4,360	.67	7.55	1.11
Europe	3,100	.69	7.58	1.08
Latin America	1,260	.63	7.47	1.17
		Learning & self-development (23 items)		
Europe and Latin America	3,892	.73	7.66	1.21
Europe	2,731	.73	7.65	1.17
Latin America	1,161	.75	7.68	1.30
		Working well with others (20 items)		
Europe and Latin America	4,185	.76	7.85	1.15
Europe	3,200	.77	7.85	1.12
Latin America	985	.73	7.82	1.23

an internship; 16% of all participants were graduate students. Students studied across thirteen different majors (31 % Business & Management, 15 % Engineering and Information Technology, 14% Health Sciences).

Results

Internal Consistency Reliabilities

We based our analyses on SJT scenario scores as sum scores for the best and worst choice per scenario. To calculate internal consistencies for each of the five SJTs, we used the full information maximum likelihood procedure and the ML estimator in Mplus Version 7.4 (Muthén & Muthén, 1998-2015) to estimate scenario scores from missing values. Then, we used intercorrelations between scenario scores to calculate Cronbach's alpha for our total sample. Internal consistencies of the five SJTs were moderate to acceptable for the total sample (.67-.78, see Table 1). Internal consistencies calculated separately for each region produced similar results (see Table 1).

Measurement Invariance across Regions

To examine measurement invariance across regions for each of the five SJTs, we first sought to establish a baseline model for the total sample, then investigated model fit for the baseline model within each region, and afterwards ran increasingly restrictive multigroup confirmatory factor analyses (e.g., Byrne & Stewart, 2006; Byrne & Van de Vijver, 2010). We conducted these analyses in Mplus via the full information maximum likelihood procedure and the ML estimator.

To guide the examination of a baseline model for the total sample, we hypothesized that a one-factor model would explain scenario scores for each SJT. This hypothesis was based upon the fact that all scenarios and response options for a specific SJT were developed to tap into one respective competency. For all five SJTs, a one-factor model showed good model fit (see Table 2). Thus, a one-factor model was chosen as baseline model in all of the following steps.

We then investigated model fit for this baseline model per region. For the SJTs of "achieving objectives" as well as "analyzing and solving problems", model fit for the baseline model within each region were at least acceptable. For the three remaining SJTs, the CFI value for the model fit within Latin America fell below the limit of acceptable model fit. Previous studies that investigated the factor structure of SJTs frequently found similar patterns and usually failed to find good model fit (with acceptable CFI values). To analyze measurement invariance, these studies then used the best fitting model as baseline model for the multigroup confirmatory factor analyses (e.g., Krumm et al., 2015; Lievens, Sackett, Dahlke, Oostrom, & De Soete, in press). In line with this approach, we kept the one-factor model as baseline model for our measurement invariance analyses.

To investigate measurement invariance, we sought to find evidence for configural and metric invariance for the baseline model across regions (see summary of Byrne & Van de Vijver, 2010). To investigate configural measurement invariance, we restricted the number of latent factors and the number of factor loadings to be equal across both regional groups. Configural measurement invariance therefore indicates that the same factorial structure explains the observed scores across regional groups. Second, we restricted the size of factor loadings to be equal across both regional groups to investigate metric measurement invariance. Metric measurement invariance thus suggests that observed scores are equally related to the assumed latent factor(s). In other words, metric measurement invariance indicates that the observed scores measure the latent factor(s) equally across (regional) groups (see, for example, Byrne & Stewart, 2006; Byrne & van de Vijver, 2010).

To examine configural and metric measurement invariance, we inspected model fit, and conducted nested model comparisons by using the chi-square difference test as well as the criterion proposed by Cheung and Rensvold (2002). These authors stated that measurement equivalence could be defended in practical terms, if increasingly restrictive confirmatory factor analyses are associated with only marginal drops in CFI values (ΔCFI < .01; see also Byrne & Stewart, 2006). With the exception of the SJT for "achieving objectives", chi-square difference tests were not significant for all five SJTs, which provides evidence for metric measurement invariance. In addition, drops in CFI values were marginal for all five SJTs (ΔCFI ≤ .008). Thus, we concluded that metric measurement equivalence could be established for all five SJTs (see Table 3). Importantly, this means that at a practical level differences in manifest mean scenario scores across regions can be compared.

Table 2. Goodness-of-fit Indices for Factor Structure Models (Overall Sample and Within Regions)

	n	$\chi^2(df)$	χ^2/df	CFI	RMSEA (90% CI)	SRMR
			Achieving objectives			
Europe and Latin America	3,666	293.63(152)**	1.93	.908	.016 (.013-.019)	.047
Europe	2,666	294.67(152)**	1.94	.885	.019 (.016-.022)	.055
Latin America	1,000	199.15(152)**	1.31	.872	.018 (.010-.024)	.081
			Adapting to change			
Europe and Latin America	4,511	254.00(170)**	1.49	.921	.010 (.008-.013)	.041
Europe	3,586	264.45(170)**	1.56	.896	.012 (.009-.015)	.046
Latin America	925	229.27(170)**	1.35	.756	.019 (.012-.026)	.093
			Analyzing and solving problems			
Europe and Latin America	4,360	240.67(152)**	1.58	.907	.012 (.009-.014)	.042
Europe	3,100	211.46(152)**	1.39	.923	.011 (.007-.015)	.045
Latin America	1,260	175.53(152)	1.15	.875	.011 (.000-.018)	.071
			Learning & self-development			
Europe and Latin America	3,892	305.26(230)**	1.33	.908	.009 (.006-.012)	.051
Europe	2,731	288.66(230)**	1.26	.901	.010 (.006-.013)	.058
Latin America	1,161	315.02(230)**	1.37	.706	.018 (.013-.023)	.100
			Working well with others			
Europe and Latin America	4,185	240.54(170)**	1.41	.949	.010 (.007-.013)	.040
Europe	3,200	209.81(170)*	1.23	.964	.009 (.004-.012)	.043
Latin America	985	273.76(170)**	1.61	.726	.025 (.019-.030)	.092

*p < .05, **p < .01.

Table 3. Tests of Measurement Invariance for One-Factor Model Underlying SJT Scores Across Participants from Europe and Latin America

Model	$\chi^2(df)$	χ^2/df	$\Delta\chi^2$	Δdf	CFI	ΔCFI	RMSEA (90% CI)	SRMR
Achieving objectives								
Equal number of factors	493.81(304)**	1.62			.882		.018 (.015-.021)	.063
Equal factor loadings	523.03(322)**	1.62	29.21*	18	.875	.007	.018 (.016-.021)	.068
Adapting to change								
Equal number of factors	493.72(340)**	1.45			.866		.014 (.011-.017)	.059
Equal factor loadings	509.51(359)**	1.42	15.79	19	.869	.003	.014 (.011-.016)	.061
Analyzing and solving problems								
Equal number of factors	386.99(304)**	1.27			.913		.011 (.007-.014)	.054
Equal factor loadings	409.39(322)**	1.27	22.40	18	.909	.004	.011 (.008-.014)	.056
Learning and self-development								
Equal number of factors	603.68(460)**	1.31			.837		.013 (.010-.015)	.073
Equal factor loadings	632.84(482)**	1.31	29.16	22	.829	.008	.013 (.010-.015)	.077
Working well with others								
Equal number of factors	483.56(340)**	1.42			.903		.014 (.011-.017)	.058
Equal factor loadings	507.38(359)**	1.41	23.82	19	.900	.003	.014 (.011-.017)	.062

*$p < .05$, **$p < .01$.

Discussion

Many educational and (non)profit organizations have investigated which skills or competencies are needed to face the challenges of the 21st century (Binkley et al., 2012; Geisinger, 2016). Subsequently, researchers have started to investigate how such 21st century skills can be best measured (Kyllonen, 2012). One such key challenge deals with assessing 21st century skills without biases that may interfere with comparing results obtained across various geographical regions and cultures. This study advances our knowledge about appropriate assessment approaches for 21st century skills by outlining how the combined emic-etic approach enables developing SJTs that tap into 21st century skills across regional groups. To this end, we investigated measurement invariance across Europe and Latin America for five different SJTs that assessed a core competency for graduating students to be successful in entry-level jobs.

Our results showed that configural and metric measurement invariance could be established across Europe and Latin America for all of the five SJTs. Thus, the same factorial structure explained SJT scenario scores across these regional groups and SJT scenario scores measured the latent factor(s) equally across those regional groups (see, for example, Byrne & Stewart, 2006; Byrne & van de Vijver, 2010). In other words, participants from Europe and Latin America interpreted the SJT scenarios and response options in the same way and attributed the same meaning to them. This is a fundamental precondition to rule out measurement effects and to investigate mean differences across (regional) groups (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000).

Our results advance knowledge about the use of SJTs across geographical regions and cultures. Given SJTs' highly contextualized nature, comparing SJT scores across regions and cultures is viewed as a crucial challenge (e.g., Lievens, 2006; Ployhart & Weekley, 2006). Previous cross-cultural investigations of SJTs also showed mixed results when the SJT development followed an imposed etic approach and did not include cross-regional/cultural input across all steps of SJT development (Lievens, Corstjens, et al., 2015; Such & Schmidt, 2004). However, as we demonstrated, integrating subject matter experts from different regions and cultures during the definition of the construct of measurement, the sampling of critical incidents, scenario writing, generation of response options, and setting the scoring key provides the fundament for SJTs to work well and be transportable across regions/cultures.

Although a combined emic-etic approach is time and resource intensive, it seems to pay off in terms of the cross-cultural application of assessment methods. Our work therefore attests to the success of relying on a combined emic-etic approach and extends similarly positive findings from research on the cross-cultural transportability of personality inventories (Cheung, Cheung et al., 2008; Cheung, Fan et al., 2008; Cheung et al., 1996; Schmit et al., 2000). To the best of our knowledge, this study is the first to apply a combined etic-emic approach of SJT development and to investigate its effects on measurement invariance across geographical regions. Our general recommendation is that the combined emic-etic approach serves as a viable strategy to develop SJTs for assessing 21st century skills across geographical regions.

Some caveats are in order, though. First, traditional, written SJTs with close-ended response formats do not measure behavior related to 21st century skills. Instead, they capture people's procedural knowledge about engaging in behavior related to these skills (Lievens, 2017; Lievens & Motowidlo, 2016; Motowidlo & Beier, 2010; Motowidlo et al., 2006). Recent research explored SJTs with other stimulus and response formats such as constructed response multimedia tests. These tests present short video clip situations to participants, that then have to display their behavioral response in front of a webcam. Evaluations of these constructed responses have been shown to be valid indicators of job and training performance (Cucina et al., 2015; De Soete, Lievens, Oostrom, & Westerveld, 2013; Herde & Lievens, 2018; Lievens, De Corte, & Westerveld, 2015; Lievens & Sackett, 2017; Lievens et al., in press; Oostrom, Born, Serlie, & van der Molen, 2010, 2011). Although constructed response multimedia tests add costs to SJT development (i.e., design of video clips and evaluation of participants' behavioral responses), they might complement current approaches to the assessment of 21st century skills. Given their dynamic audiovisual stimulus format and their audiovisual constructed response format, constructed response multimedia tests are even more contextualized than written, close-ended SJTs. Future research should therefore investigate whether constructed response multimedia tests developed according to a combined emic-etic approach also produce scores of 21st century skills that can be compared across regions and cultures.

As another limitation, we had data for only two geographical regions (Europe and Latin America). That said, this sample incorporated participants from eighteen different countries, thereby attesting to a huge cultural diversity. Nonetheless, further empirical research is necessary to replicate our results and examine the comparability of scores derived from SJTs across other geographical regions and cultures.

Conclusion

In sum, this paper is the first to investigate the combined emic-etic approach to develop SJTs to obtain scores that can be compared across geographical regions and cultures. Our results established metric measurement invariance across five SJTs for participants from Europe and Latin America. Hence, this study attests to the potential of the combined emic-etic approach. We therefore encourage researchers and practitioners to adopt this approach in cross-cultural research and practice for assessing 21st century skills.

Conflict of Interest

The authors of this article declare no conflict of interest.

Acknowledgements

We thank Jonas W. B. Lang and Bert Weijters as well as Linda and Bengt Muthén for helpful advice on our data analyses.

Note

[1]We re-ran our analyses once with only students included who did the SJTs in their dominant language and once only with students included who reported to be "very comfortable" with the test language. Given that results were similar and did not change conclusions, we report results for our complete sample only.

References

Ainley, J., Fraillon, J., Schulz, W., & Gebhardt, E. (2016). Conceptualizing and measuring computer and information literacy in cross-national contexts. *Applied Measurement in Education, 29*, 291-309. https://doi.org/10.1080/08957347.2016.1209205

Bartram, D. (2012). *The SHL universal competency framework* (SHL White Paper). Thames Ditton, UK: SHL Group.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4*, 119-128. https://doi.org/10.1080/00207596908247261

Berry, J. W. (1990). Imposed etics, emics, and derived etics: Their conceptual and operational status in cross-cultural psychology. In T. N. Headland, K. L. Pike, & M. Harris (Eds.), *Frontiers of anthropology, Vol. 7. Emics and etics: The insider/outsider debate* (pp. 84-99). Thousand Oaks, CA: Sage Publications.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17-66). Dordrecht, Nederland: Springer. https://doi.org/10.1007/978-94-007-2324-5_2

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71,* 460-502. https://doi.org/10.1177/0013164410375112

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13,* 287-321. https://doi.org/10.1207/s15328007sem1302_7

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10,* 107-132. https://doi.org/10.1080/15305051003637306

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B. (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64,* 225-262. https://doi.org/10.1111/j.1744-6570.2010.01207.x

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of collaborative problem solving in education environments. *Applied Measurement in Education, 29*, 250-264. https://doi.org/10.1080/08957347.2016.1209204

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

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233-254. https://doi.org/10.1207/S15327043HUP1503_01

Cheung, F. M., Cheung, S. F., Jianxin Zhang, Leung, K., Leong, F., & Kuang Huiyeh. (2008). Relevance of openness as a personality dimension in Chinese culture: Aspects of its cultural relevance. *Journal of Cross-Cultural Psychology, 39*, 81-108. https://doi.org/10.1177/0022022107311968

Cheung, F. M., Fan, W., Cheung, S. F., & Leung, K. (2008). Standardization of the cross-cultural Chinese Personality Assessment Inventory for adolescents in Hong Kong: A combined emic-etic approach to personality assessment. *Acta Psychologica Sinica, 40*, 839-852. https://doi.org/10.3724/SP.J.1041.2008.01639

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese personality assessment inventory. *Journal of Cross-Cultural Psychology, 27,* 181-199. https://doi.org/10.1177/0022022196272003

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. https://doi.org/10.1207/S15328007SEM0902_5

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

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69*, 979-1006. https://doi.org/10.1111/1467-6494.696172

Church, A. T., & Lonner, W. J. (1998). The cross-cultural perspective in the study of personality: Rationale and current research. *Journal of Cross-Cultural Psychology, 29*, 32-62. https://doi.org/10.1177/0022022198291003

Cucina, J. M., Su, C., Busciglio, H. H., Harris Thomas, P., & Thompson Peyton, S. (2015). Video-based testing: A high-fidelity job simulation that demonstrates reliability, validity, and utility. *International Journal of Selection and Assessment, 23,* 197-209. https://doi.org/10.1111/ijsa.12108

De Soete, B., Lievens, F., Oostrom, J., & Westerveld, L. (2013). Alternative predictors for dealing with the diversity–validity dilemma in personnel selection: The constructed response multimedia test. *International Journal of Selection and Assessment, 21,* 239-250. https://doi.org/10.1111/ijsa.12034

Ercikan, K., & Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education, 29,* 310-318. https://doi.org/10.1080/08957347.2016.1209210

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358. https://doi.org/10.1037/h0061470

Geisinger, K. F. (2016). 21st century skills: What are they and how do we assess them? *Applied Measurement in Education, 29*, 245-249. https://doi.org/10.1080/08957347.2016.1209207

Getha-Taylor, H., Hummert, R., Nalbandian, J., & Silvia, C. (2013). Competency model design and assessment: Findings and future directions. *Journal of Public Affairs Education, 19*, 141-171. https://doi.org/10.1080/15236803.2013.12001724

Greiff, S., & Kyllonen, P. (2016). Contemporary assessment challenges: The measurement of 21st century skills. *Applied Measurement in Education, 29,* 243-244. https://doi.org/10.1080/08957347.2016.1209209

Headland, T. N., Pike, K. L., & Harris, M. (1990). *Frontiers of anthropology, Vol. 7. Emics and etics: The insider/outsider debate.* Thousand Oaks, CA: Sage Publications.

Heine, S. J., & Buchtel, E. E. (2009). Personality: The universal and the culturally specific. *Annual Review of Psychology, 60*, 369-394. https://doi.org/10.1146/annurev.psych.60.110707.163655

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology, 82*, 903-918. https://doi.org/10.1037//0022-3514.82.6.903

Herde, C. N., & Lievens, F. (2018). Multiple speed assessments: Theory, practice, & research evidence. *European Journal of Psychological Assessment.* Advance online article. https://doi.org/10.1027/1015-5759/a000512

Herde, C. N., Wüstenberg, S., & Greiff, S. (2016). Assessment of complex problem solving: What we know and what we don't know. *Applied Measurement in Education, 29,* 265-277. https://doi.org/10.1080/08957347.2016.1209208

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly, 74,* 201-222. https://doi.org/10.1093/poq/nfq011

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*, 296-309. https://doi.org/10.1177/0022022189203004

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1,* 93-114. https://doi.org/10.1207/S15327574IJT0102_1

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36,* 264-277. https://doi.org/10.1177/0022022104272905

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

Kurz, R., & Bartram, D. (2002). Competency and individual performance: Modelling the world of work. In I. T. Robertson, M. Callinan, & D. Bartram (Eds.), *Organizational effectiveness: The role of psychology* (pp. 227-255). Chichester, UK: Wiley.

Kyllonen, P. C. (2012). *Measurement of 21st century skills within the common core state standards*. Presented at the Invitational Research Symposium on Technology Enhanced Assessments (TEA). Washington, DC. Retrieved from https://www.ets.org/Media/Research/pdf/session5-kyllonen-paper-tea2012.pdf

Lee, Y. (2009). Competencies needed by Korean HRD master's graduates: A comparison between the ASTD WLP competency model and the Korean study. *Human Resource Development Quarterly, 20*, 107-133. https://doi.org/10.1002/hrdq.20010

Leong, F. T. L., Leung, K., & Cheung, F. M. (2010). Integrating cross-cultural psychology research methods into ethnic minority psychology. *Cultural Diversity and Ethnic Minority Psychology, 16*, 590-597. https://doi.org/10.1037/a0020127

Lievens, F. (2006). International situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 279-300). Mahwah, NJ: Erlbaum.

Lievens, F. (2017). Construct-driven SJTs: Toward an agenda for future research. *International Journal of Testing, 17*, 269-276. https://doi.org/10.1080/15305058.2017.1309857

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*, 361-372. https://doi.org/10.1111/ijsa.12120

Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures: Effects of response fidelity on performance and validity. *Journal of Management, 41*, 1604-1627. https://doi.org/10.1177/0149206312463941

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9*, 3-22. https://doi.org/10.1017/iop.2015.71

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

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (in press). Constructed response formats and their effects on minority–majority differences and validity. *Journal of Applied Psychology.* https://doi.org/10.1037/apl0000367

Lucas, B. (2016). A five-dimensional model of creativity and its assessment in schools. *Applied Measurement in Education, 29*, 278-290. https://doi.org/10.1080/08957347.2016.1209206

Lunev, A., Petrova, I., & Zaripova, V. (2013). Competency-based models of learning for engineers: A comparison. *European Journal of Engineering Education, 38*, 543-555. https://doi.org/10.1080/03043797.2013.824410

Lytle, A. L., Brett, J. M., Barsness, Z. I., Tinsley, C. H., & Janssens, M. (1995). A paradigm for confirmatory cross-cultural research in organizational behavior. *Research in Organizational Behavior, 17*, 167-214.

McDaniel, M. A., Hartman, N., Whetzel, D. L., & Grubb III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

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

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. G. Rumsey & C. B. Walker (Eds.), *Personnel selection and classification* (pp. 193-232). Hillsdale, NJ: Erlbaum.

Morris, M. W., Leung, K., Ames, D., & Lickel, B. (1999). Views from inside and outside: Integrating emic and etic insights about culture and justice judgment. *Academy of Management Review, 24*, 781-796. https://doi.org/10.5465/amr.1999.2553253

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. https://doi.org/10.1037/a0017975

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

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. https://doi.org/10.1037/0021-9010.91.4.749

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (7th Edition). Los Angeles, CA: Muthén & Muthén.

Nishii, L. H., Ployhart, R. E., Sacco, J. M., Wiechmann, D., & Rogg, K. L. (2001). *The influence of culture on situational judgment test responses.* Presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology. San Diego, CA.

OECD. (2014). *PISA 2012 Technical Report*. Paris, France: OECD Publishing.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology, 19*, 532-550. https://doi.org/10.1080/13594320903000005

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology, 10*, 78-88. https://doi.org/10.1027/1866-5888/a000035

Pike, K. L. (1967). Etic and emic standpoints for the description of behavior. In K. L. Pike (Ed.), *Language in relation to a unified theory of the structure of human behavior* (pp. 37-72). Den Haag, Nederland: Mouton & Co.

Ployhart, R. E., & Weekley, J. A. (2006). Situational judgment: Some suggestions for future science and practice. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 345-350). Mahwah, NJ: Erlbaum.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment test comparable? *Personnel Psychology, 56*, 733-752. https://doi.org/10.1111/j.1744-6570.2003.tb00757.x

Sahoo, F. M. (1993). Indigenisation of psychological measurement: Parameters and operationalisation. *Psychology and Developing Societies, 5*, 1-13. https://doi.org/10.1177/097133369300500101

Schmit, M. J., Kihm, J. A., & Robie, C. (2000). Development of a global measure of personality. *Personnel Psychology, 53*, 153-193. https://doi.org/10.1111/j.1744-6570.2000.tb00198.x.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497-508. https://doi.org/10.1037/0021-9010.89.3.497

Strong, M. H., Burkholder, G. J., Solberg, E. G., Stellmack, A., & Presson, W. D. (2019). *Development and validation of a global competency framework for preparing new graduates for early career professional roles.* Manuscript submitted for publication.

Such, M. J., & Schmidt, D. B. (2004). *Examining the effectiveness of empirical keying: A cross-cultural perspective.* Presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology. Chicago, IL.

Van de Vijver, F. J. R. (2003). Test adaptation/translation methods. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 960-964). Thousand Oaks, CA: Sage Publications.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. https://doi.org/10.1177/109442810031002

Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior, 2,* 295-322. https://doi.org/10.1146/annurev-orgpsych-032414-111304

Weekley, J. A., 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 judgement tests: Theory, measurement, and application* (pp. 157-182). Mahwah, NJ: Lawrence Erlbaum.

Yang, K.-S. (2000). Monocultural and cross-cultural indigenous approaches: The royal road to the development of a balanced global psychology. *Asian Journal of Social Psychology, 3*, 241-263. https://doi.org/10.1111/1467-839X.00067

Appendix

Definitions of SJT Competencies

Competency	*Definition*
Achieving objectives	Accepts or sets demanding individual goals. Meets individual goals and objectives. Takes initiative to seek additional responsibilities, as appropriate. Evaluates work outcomes to ensure quality standards are met.
Adapting to change	Adjusts work style and interpersonal behavior to fit different situations and environments. Accepts and integrates new ideas and information on their merits. Supports and complies with change initiatives. Works effectively when faced with ambiguity.
Analyzing & solving problems	Critically evaluates information and its sources. Identifies gaps in information and seeks appropriate sources to close them. Synthesizes and integrates information into what is already known about a topic. Recognizes patterns in information to identify the bigger picture. Follows best practices and appropriately analyzes quantitative and qualitative data. Identifies and independently solves work problems, as appropriate. Considers multiple approaches when solving problems.
Learning & self-development	Identifies and addresses own knowledge gaps and training needs. Continually expands own knowledge and skills. Applies knowledge and training to professional contexts. Critically evaluates own strengths and weaknesses and pursues development. Seeks feedback and learns from successes and failures. Learns from others and seeks mentors.
Working well with others	Develops and maintains effective working relationships. Interacts effectively with people from different backgrounds. Listens to others and values and incorporates diverse viewpoints. Supports team decisions once they have been made. Adjusts own workload to help meet team commitments, as appropriate. Recognizes and demonstrates empathy for others' feelings, needs, and concerns. Appropriately resolves own work disagreements.