

13

International Situational Judgment Tests

Filip Lievens
Ghent University, Belgium

Your sports club is planning a trip to Berlin to attend the Germany-England football game, which will take place in 2 weeks. You have been entrusted with the preparations and entire management of the trip. What do you intend to do?

(—Ansbacher (1941, p. 381, cited in Highhouse, 2002)

The above item was given in the late 1930s to German employees to measure something else other than cognitive ability (e.g., planning and organizing). It illustrates that situational judgment tests (SJTs) have been used outside the United States for quite some time. Early international applications of SJTs can also be found in the so-called cultural assimilators in cross-cultural training programs (Bhawuk & Brislin, 2000). In these cultural assimilators, future expatriates are presented with written situations of an expatriate interacting with a host national and are asked to indicate the most effective response alternative.

Recently, there has been a renewed interest in the use of SJTs in an international selection context. One of the key reasons is that the globalization of the economy necessitates organizations to view the labor market in an international scope and to develop selection procedures that can be used across multiple countries. However, there is a dearth of research on international selection in general and on international SJTs in particular. This leaves many questions unanswered. Most importantly, a key issue is

whether SJTs developed in one culture can be transported to and used as a valid predictor in another culture?

The purpose of this chapter is to explore the limits of the generalizability of SJTs and their criterion-related validity across cultural boundaries. The chapter begins with a brief review of personnel selection in an international context. Next, I focus on the criterion-related validity of SJTs across cultures. Specifically, I delineate the factors under which the criterion-related validity of SJTs might generalize to foreign countries and across-country applications. Finally, I offer insights into best practices.

A BRIEF OVERVIEW OF INTERNATIONAL SELECTION RESEARCH

Despite the growing importance of selection in an international context, there is a paucity of internationally oriented selection research (see Lievens, *in press*, for a review). Most prior studies were descriptive in nature and compared selection practices from one country to another. Generally, considerable variability in selection procedure usage across countries was found (see Newell & Tansley, 2001; Shackleton & Newell, 1997). More recent studies have started to examine why selection procedures are used differentially across countries. In particular, the multi-country study of Ryan, McFarland, Baron, and Page (1999) found some evidence that one of Hofstede's (1991) dimensions (i.e., uncertainty avoidance) could explain some of the variability in selection practices. For example, organizations in cultures high in uncertainty avoidance used more selection methods, used them more extensively, and conducted more interviews.

Other internationally oriented selection studies focused on the perceptions of selection procedures across different countries. Steiner and Gilliland (2001) reviewed these studies and concluded that a fairly consistent picture emerged as the same selection procedures (interviews, resumes, and work samples) received favorable reactions across various countries. In all countries, job-relatedness also emerged as the key determinant of favorable perceptions. Steiner and Gilliland (2001) explained these similarities on the basis of the shared European heritage of the countries reviewed (Belgium, France, Spain, South Africa, and the United States).

Finally, a limited amount of studies has tackled the fundamental question as to whether the criterion-related validity of a selection procedure will differ when used in other countries and cultures. Essentially, two hypotheses have been proposed, namely the validity generalization hypothesis and the situational specificity hypothesis (Salgado & Anderson, 2002). The validity generalization hypothesis states that observed criterion-related validity coefficients will vary because of statistical artifacts (such

13. INTERNATIONAL SJTs**281**

as sampling error, range restriction, criterion unreliability). When these statistical artifacts are accounted for, criterion-related validity coefficients will generalize across different situations (jobs, occupational groups, organizations; F. Schmidt & Hunter, 1984). In an international context, this means that criterion-related validity coefficients associated with a specific selection procedure obtained in one country will generalize to another country. Exactly the opposite is posited by the situational specificity hypothesis. According to this hypothesis, there should be high variability in the observed criterion-related validity coefficients obtained in different situations (jobs, occupational groups, organizations). Whenever the situation changes, the observed criterion-related validity coefficient might also change (F. Schmidt & Hunter, 1984). Applied to an international context, this means that selection procedures might be valid in one country but not in another country.

Few empirical studies have tested these competing hypotheses. To our knowledge, only the generalizability of the criterion-related validity of cognitive-ability tests and personality inventories has been put to the test in an international context. Generally, results have provided support for the validity generalization hypothesis. For example, Salgado Anderson, Moscoso, Bertua, and De Fruyt (2003), and Salgado, Anderson, et al., (2003) found evidence for validity generalization for cognitive ability tests across seven European countries. In addition, the magnitude of the criterion-related validity coefficients found conformed to previous U.S. meta-analyses (F. Schmidt & Hunter, 1998), underscoring that cognitive-ability validities generalized across jobs, occupations, and borders. In the domain of personality tests, fairly consistent results have also been found across American (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991) and European (Salgado, 1997) meta-analyses, with mainly Conscientiousness emerging as a consistent predictor across jobs, occupations, and cultures.

THE CRITERION-RELATED VALIDITY OF SJTs ACROSS CULTURES

Our overview of prior international selection research showed that the criterion-related validity of cognitive ability tests and personality inventories generalized across countries. Does this mean that the same consistent results in terms of criterion-related validity will be found for SJTs across countries and cultures? Equally important, which factors can be expected to impact on the potential generalizability of SJTs across cultures? The remainder of this chapter focuses on these issues. Specifically, I discuss three influencing factors: (a) the cultural transportability of SJT item

characteristics, (b) the point-to-point correspondence between predictor and criterion, and (c) the type of constructs measured by SJTs.

The Cultural Transportability of SJT Item Characteristics

SJT Item Characteristics. SJT items are highly contextualized because they are embedded in a particular context or situation. 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 (Cropanzano, 1998; Lytle, Brett, Barsness, Tinsely, & Janssens, 1995). Table 13.1, presents a matrix in which SJT item characteristics are cast in terms of Hofstede's (1991) cultural dimensions (i.e., individualism/collectivism, power distance, uncertainty avoidance, and masculinity/femininity). Note that I use Hofstede's (1991) framework, although I acknowledge it is also possible to construct this matrix with other cultural frameworks (House, Hanges, Javidan, Dorfman, & Gupta, 2004; Kluckhohn & Strodtbeck, 1961; Schwartz & Bardi, 2001; Schwartz & Sagiv, 1995; Trompenaars & Hampden-Turner, 1997).

The problem situations (as reflected in the item stems of SJTs) that are presented to candidates in a written or video-based format are a first characteristic of SJTs. These problem situations are generated from a job analysis and from critical incidents provided by high and low performers on a specific criterion (job). When SJTs are used in an international context, the issue then becomes whether there are cultural differences in terms of the situations (critical incidents) generated (holding the type of job constant). According to value orientations theory (Kluckhohn & Strodtbeck, 1961), all cultures encounter very common problem situations. Applied to SJTs, this would mean that for a given job the situations encountered might be fairly similar across various cultures. However, we believe a lot also depends on the type of situations studied. In cross-cultural psychology, generalizability has typically been studied and found for major situations such as situations of joy, fear, anger, sadness, disgust, shame, and guilt (Scherer & Wallbott, 1994; Scherer, Wallbott, & Summerfield, 1986). Such basic configurations are different from many of the work-related situations included in SJTs. Therefore, we do not expect the situations and response options of SJTs to generalize across cultures. Some situations will simply not be relevant in one culture, whereas they might be very relevant in another culture. Think about the differences in organizing meetings across countries. For example, applicants in a culture high on power distance might have trouble answering a situation about a party off hours between employees and their boss. If one does not take account of these cultural differences, it might well be that applicants are presented with an SJT item stem that is simply

13. INTERNATIONAL SJTs**283**

not relevant in their culture. To our knowledge, no empirical studies have tested whether similar situations are generated across cultures.

A second item characteristic of SJTs are the response alternatives. These are taken from possible ways of dealing with a given situation as provided by high and low performers on a specific criterion (job). In an international context, one might question whether the response alternatives given to applicants are transportable from one culture to another. Value orientations theory (Kluckhohn & Strodtbeck, 1961) posits that all cultures encounter similar problems and that all cultures discovered similar responses to these problems. Yet, the same argument as earlier applies here. The possible responses to the basic configurations studied in cross-cultural psychology do not echo the specific responses to the work-related situations depicted in SJTs. Thus, we expect that the possible range of relevant response options might differ from one culture (holding the type of situations and the type of job constant) to another. If one does not take account of these differences, the SJT might present applicants with response options that are not relevant in a given culture. This also means that the response endorsement frequencies might differ from one culture to another. So, what might be a good distractor (e.g., yelling in a meeting when no one is listening to your opinion) in one culture (e.g., culture low in power distance) might not be endorsed by many applicants in another (e.g., culture high in power distance).

Apart from item stems and response alternatives, the SJT *scoring key* is a third component of all SJTs (McDaniel & Nguyen, 2001). The correct answer on an SJT is determined either empirically (by comparing low and high performers) or rationally (by experts), although a hybrid of these two approaches is sometimes followed. It is expected that cultural differences will affect the effectiveness of response options and therefore the scoring key of SJTs. This expectation is based on value orientations theory (Kluckhohn & Strodtbeck, 1961) and attribution theory (Bond, 1983; Morris & Peng, 1994).

According to value orientations theory (Kluckhohn & Strodtbeck, 1961), cultures differ in terms of their preference for specific responses to problem situations. This is illustrated by linking Hofstede's dimensions with the effectiveness of response alternatives to SJT items. For instance, in a culture high on uncertainty avoidance, the effective response to a specific written SJT situation (e.g., supervising a group of young employees) might be to impose rules and structure. However, the same reply to the same situation might be valued as ineffective in a culture low on uncertainty avoidance because ambiguity is not perceived as a threat. The individualism–collectivism might also affect SJT response effectiveness. In fact, we expect that answers that promote group harmony might be considered more

effective in cultures high in collectivism, whereas the reverse might be true in cultures low on individualism. The masculinity–femininity dimension might affect SJT responses, such that answers that involve competition might be preferred in cultures high on masculinity. Finally, answers that minimize ambiguity and appear decisive might be considered most effective in a culture high on uncertainty avoidance.

Attribution theory also posits that the effectiveness of responses to situations might differ from one culture to another. This is because attribution patterns reflect implicit theories acquired from socialization in a specific culture. Therefore, they are differentially distributed across human cultures (Bond, 1983; Morris & Peng, 1994). For example, Morris and Peng's study revealed that American people attributed social events more to personal dispositions (i.e., attributions based on the belief that social behavior expresses stable, global, and internal dispositions), whereas Chinese people attributed more to situational factors (attributions based on the belief that social behavior is shaped by relationships, roles, and situational pressures). The evidence that attribution patterns are differentially distributed across human cultures serves as the foundation of the so-called cultural assimilators that are often used in cross-cultural training (Bhawuk & Brislin, 2000). Cultural assimilators share similarities with SJTs because they also present written or video-based situations to individuals. A difference is that the situation is always a social situation in another culture and that the response alternatives given are essentially possible attributions associated with the event depicted. According to Bhawuk and Brislin, cultural assimilators aim to teach expatriates to make isomorphic attributions. This means that individuals attribute a social event in a specific culture in the same way as is done in that specific culture.

A fourth item characteristic that might be prone to cultural differences is the link between response options as indicators for a given construct. Unlike cognitive-ability tests, we expect that the item-construct relationship in SJTs is more susceptible to deficiency and contamination because of possible cross-cultural differences in the meaning/interpretation of the same situation content or same response to the same situation. For example, given the same written situation (e.g., a situation depicting a meeting between an older supervisor and a group of employees), the same behavior (e.g., clearly and openly defending one's views about work standards in front of the supervisor with all employees being present) might be linked to a specific construct (e.g., assertiveness) in one culture (culture low in power distance), whereas it might be an indicator for another construct (e.g., rudeness, impoliteness) in another culture (culture high in power distance).

13. INTERNATIONAL SJTs**285**

Empirical Research. Studies that have examined the cultural transportability of SJT item characteristics are very scarce. As noted, no studies have explored cultural differences in terms of the situations, response options, or response option–construct linkages. We retrieved only one empirical study that examined whether the preference (effectiveness) for response alternatives differs across cultures. Nishii, Ployhart, Sacco, Wiechmann, and Rogg (2001) conducted a study among incumbents of a multinational food chain in different countries (Canada, Germany, Korea, Mexico, Spain, the United Kingdom, and Thailand). They investigated whether the endorsement of response options to five SJT items was affected by culture. Cultural dimensions were operationalized in terms of Hofstede's cultural dimensions. Results revealed that people of different cultural backgrounds were differentially attracted to specific response alternatives, and that these differences were consistent with theoretical expectations. As a matter of fact, people from individualistic cultures chose response options that were task-oriented and that involved communicating directly with others. However, for the same item, people from collectivistic cultures tended to choose response options with a focus on group harmony and protecting others' face.

On a more general level, a wide variety of empirical research in cross-cultural psychology (e.g., Smith, Dugan, Peterson, & Leung, 1998; Smith et al., 2002) has also shown that the effectiveness of work-related behaviors in response to a given situation might drastically differ across cultures. As there are numerous examples are, we cite only two studies. Adler, Doktor, and Reddin (1986) showed that there were differences in decision making and information processing across cultures and countries. As an example, they mentioned that Japanese people like to go from general to specific, whereas Western people prefer to get rid of details before talking about larger issues. S. Schmidt and Yeh (1992) drew similar conclusions with regard to differences in leadership behaviors and styles across cultures.

The Point-to-Point Correspondence Between Predictor and Criterion

SJTs as Externally Constructed Measures. SJTs are fundamentally different measures than cognitive-ability or personality tests. Cognitive-ability tests and to a certain extent also personality inventories are internally constructed predictor measures (Mount, Witt, & Barrick, 2000). These predictor measures are typically decontextualized and are developed to have generalizability across a wide variety of situations. Accordingly, it is expected that the criterion-related validity of these measures

will generalize across jobs, occupations, and cultures. As noted earlier, this expectation has been confirmed so far.

Conversely, contextualized measures such as SJTs are externally constructed because they are also developed for a very specific criterion. In fact, SJT items are directly developed or sampled from the criterion behaviors that the test is designed to predict (Chan & Schmitt, 2002). Apart from SJTs, other examples of externally constructed measures include situational interviews, behavior-description interviews, work samples, and assessment center exercises.

For externally constructed predictors such as SJTs, the point-to-point correspondence between the predictor and the criterion domain is of paramount importance as it gives them their criterion-related validity. This contrasts to internally oriented measures such as cognitive ability tests whose criterion-related validity is expected to generalize across a wide variety of jobs and occupations. When framed in this way, it should be clear that using an SJT in a different culture than originally intended is conceptually not different from using an SJT for another job or occupation than originally intended. In other words, the fact that an SJT is used in another culture does not make it invalid per se. As long as one ensures that the predictor and criterion domains match, criterion-related validity will be high. Conversely, when the predictor and criterion domains do not overlap, criterion-related validity will be low. All of this is based on the well-known notion that validity is about matching predictor and criterion domains (Binning & Barrett, 1989).

To examine these expectations, we categorized possible international applications of SJTs along these two dimensions (predictor and criterion). A further distinction is made between “national” (original culture) and “international” (host culture). As the vast majority of SJT practice and research has been conducted in the United States, we take the United States as point of reference. This means that “national” applications refer to the use of SJTs in the United States. International applications, in turn, refer to the use of SJTs outside the United States.

The “predictor–criterion” distinction and the “national–international” distinction lead to four quadrants. These quadrants are presented in Fig. 13.1. The following section discusses the main criterion-related validity issues for each of these four quadrants. When available, prior SJT criterion-related validity studies are reviewed.

Within-Culture Applications. Quadrant A of Fig. 13.1 does not really deal with SJT research in an international context because it consists of studies wherein the predictor (SJT) was developed and used in the

13. INTERNATIONAL SJTs

287

	Criterion	
	National	International
National Predictor (SJT)	A	B
	C	D
International		

FIG. 13.1. Overview of international applications of situational judgment tests.

United States. Afterward, the criterion data (job-performance data) were also gathered in the United States.

Given that most prior SJT studies have been conducted in the United States, this quadrant consists of the majority of SJT research. In fact, McDaniel, Morgeson, Finnegan, Campion, and Braverman (2001) meta-analyzed 39 prior SJT studies (that generated 102 criterion-related validity coefficients) conducted in the United States. Results showed that SJTs were valid predictors, with an estimated population validity of .34. Other studies conducted in the United States (e.g., Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt Harvey, 2001) have further shown that SJTs have incremental validity over and above cognitive ability and personality tests.

Quadrant D of Fig. 13.1 also entails within-culture applications of SJTs. In studies in Quadrant D, the predictor (SJT) was developed and used outside the United States. Afterward, the criterion data were also gathered outside the United States. Studies in Quadrant D used a so-called *emic approach* (Berry, 1969). This means that SJTs are developed and validated with the own culture as the point of reference. One example is the study of Chan and Schmitt (2002). These researchers developed an SJT for civil service positions in Singapore. Although the development of the SJT conformed to the procedures used in U.S. SJTs (see Motowidlo, Hanson, & Crafts, 1997), 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. Another example is the development of a video-based SJT for use in the Belgian admission exam “Medical and Dental Studies” (Lievens, Buyse, & Sackett, 2005; Lievens & Coetsier, 2002).

Again, the development of the SJT closely followed U.S. studies, while at the same time ensuring that the job-relevant scenarios were derived from input of local experts. Although SJTs seem to be less popular outside the United States, we found other examples of SJT studies in Quadrant D in Germany (Behrmann, 2004; Funke & Schuler, 1998; Kleinmann & Strauss, 1998; Schuler, Diemand, & Moser, 1993), the Netherlands (Born, 1994; Born, Van der Maesen de Sombreff, & Van der Zee, 2001; Van Leest & Meltzer, 1995), Korea (Lee, Choi, & Choe, 2004), and China (Jin & Wan, 2004).

Given the clear overlap between predictor and criterion contexts, we see no reason why carefully developed SJTs would not be valid in the applications mentioned in Quadrant D. Empirical research attests to this. Chan and Schmitt (2002) found that their SJT was a valid predictor for overall performance. In addition, their SJT application in Singapore had incremental validity over cognitive ability, personality, and job experience. This corresponds to the aforementioned studies in the United States. Similarly, Lievens et al. (2005) found that a video-based SJT was a valid predictor of Belgian medical students' performance on interpersonally oriented courses and had incremental validity over cognitive ability for predicting these courses. Funke and Schuler (1998) showed that their SJT was predictive for German students' performance on interpersonally oriented role-plays. Finally, Behrmann's (2004) study revealed that the initial criterion-related validity results of an SJT developed for German call center agent incumbents were promising.

Across-Culture Applications. Quadrant B of Fig. 13.1 consists of studies wherein the SJT was developed in the United States. However, it was used and validated in a different culture. Thus, contrary to Quadrants A and D, Quadrant B involves across-country applications of SJTs. The studies in Quadrant B are also examples of an imposed etic approach (Berry, 1969) as it is assumed that pre-existing assessment techniques (e.g., an SJT developed in the United States) can be adapted to different countries. For example, an SJT designed for a particular job in the United States might be used in other countries where the organization operates. Another example is the selection of people in the United States for international assignments. Once selected, these expatriates might be evaluated in the host culture (outside the United States).

Empirical research in Quadrant B is scarce. Such and Schmidt (2004) validated an SJT in four countries. The SJT and its scoring key were developed on the basis of a "cross-cultural" job analysis across multiple 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 illustrate that the

13. INTERNATIONAL SJTs**289**

criterion-related validity of an SJT might be undermined when the predictor and criterion domains do not overlap. As noted previously, given the substantial cultural differences in what is considered effective behavior in a given situation, it seems impossible to determine a universal scoring key. So, although attempts were made to ensure that the scoring key was cross-culturally oriented, we believe that the results indicate 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). Hence, the SJT was predictive only for job performance as rated in the United Kingdom and Australia but not in Mexico. In general, Nishii et al. (2001) succinctly summarized the problem as follows:

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. (p. 10)

Applications of SJTs in Quadrant C of Fig. 13.1 are even more scarce. This quadrant is comprised of applications wherein the SJT was developed outside the United States. However, it was used and validated in the United States. The selection of impatriates (people from foreign countries that are assigned to work in the corporate headquarters in the United States) on the basis of an SJT might be an example of such a cross-country application of SJTs. In a similar vein, international personnel might be selected on the basis of SJTs in a European country. Afterward, they are sent to the United States where U.S. managers evaluate them. We were not able to retrieve criterion-related validity studies of SJTs in such contexts. On the basis of the logic just explained, we expect that the criterion-related validity of the SJT will suffer in case of a lack of predictor and criterion overlap.

The difficulties related to predictor and criterion match that might be encountered in across-culture applications of SJTs are further exemplified when one takes into consideration that cultural differences might affect the criterion itself (Ployhart, Wiechmann, Schmitt, Sacco, & Rogg, 2003). In fact, in individualistic cultures, task performance is typically given more importance than contextual performance in global job performance ratings (Johnson, 2001; Rotundo & Sackett, 2002). However, it might well be that the relative importance attached to task performance vis-à-vis contextual performance when combining ratings into a summary job-performance rating might be different in other cultures (Arthur & Bennett, 1997). For example, in collectivist cultures, job-performance ratings may resemble more closely measures of contextual performance—at least as those concepts are defined

in these cultures. The key point here is that the criterion-related validity of an SJT in a host culture (country) might vary considerably depending on the relative weights given to specific performance components (task vs. contextual performance) when defining the construct of job performance in that specific culture (country) (Murphy & Shiarella, 1997).

Possible Moderators. The importance of the point-to-point correspondence between SJT and the criterion might be moderated by at least two factors. First, careful attention to matching predictor and criterion domains in international use of selection procedures might be less important for cognitively loaded SJTs than for noncognitive SJTs. As discussed here, cognitive constructs seem to be less susceptible to cultural variation. Second, the validity of cross-cultural applications is dependent on the culture specificity of the job in question. This refers to the issue as to what extent the same job-relevant behaviors on this same job are evaluated differently between cultures. If a job is not very culture-dependent/susceptible, it should not matter to validity whether SJT test development, scoring key development (expert's judgments), and criterion-domain judgments (performance ratings) were done in the same or a different culture from the culture that the test is used. Conversely, it should matter if the job is culture-dependent/susceptible. As argued by Furrer, Liu, and Sudharshan (2000), customer service quality might be an example of a job dimension that is especially susceptible to cultural differences (see also Ployhart et al., 2003).

The Type of Constructs Measured by SJTs

In recent years, an increasing amount of studies have tried to uncover which are the constructs underlying SJTs. Most studies *a posteriori* correlated SJT scores with measures of cognitive ability or personality. The meta-analysis of McDaniel et al. (2001) examined the relationship between cognitive ability and SJT scores. Correlations varied considerably (between .17 and .75), with an average correlation of .46. Another meta-analysis concentrated on noncognitive correlates of SJT scores. McDaniel and Nguyen (2001) found that SJTs correlated with most of the Big Five personality traits. Apart from cognitive ability and personality, SJTs have also been found to correlate with experience and job knowledge. Although the debate about the constructs underlying SJTs is still ongoing, there is general consensus that—similar to assessment center exercises or structured interviews—SJT are basically methods that can be designed to measure a variety of cognitive and noncognitive constructs. This notion is best exemplified by recent efforts to *a priori* build constructs into SJTs (Motowidlo, Diesch, & Jackson, 2003; Ployhart & Ryan, 2000).

13. INTERNATIONAL SJTs

291

Why might the nature of the constructs (cognitive vs. noncognitive) measured by SJTs have important implications on the generalizability of their criterion-related validity across cultures? The main reason relates to the finding that cognitive constructs are more robust to cultural variation. In fact, cognitive ability has emerged as the best stand-alone predictor whose validity generalizes across jobs, occupations, and countries (Salgado et al., 2003a, b). The key advantage of working with constructs is that it provides a basis for predicting the criterion-related validity of specific constructs measured by other methods (Hattrupp, Schmitt, & Landis, 1992; Schmitt & Chan, 1998). In particular, applied to SJTs, this would mean that SJTs that are cognitively loaded would exhibit more cross-cultural validity than SJTs that are not cognitively loaded, all other things being equal. According to Chan and Schmitt (2002), the g-loadedness of an SJT is dependent on the nature of the SJT test content. Thus, the more the SJT content is loaded with cognitive ability, the more likely it will exhibit cross cultural validity, all other things being equal.

DIRECTIONS FOR FUTURE RESEARCH

One of the common threads running through this chapter is that research on SJTs in an international context is scarce. Therefore, it was no surprise that throughout this chapter, future research needs have been suggested. In this section, I summarize these various research needs in six key directions for future research regarding the use of SJTs in an international context.

First, studies are needed that examine how culture affects the various steps in SJT development. In the matrix of Table 13.1, prior research (Nishii et al., 2001) has concentrated only on the fourth column, namely how cultural differences impact on response-choice effectiveness. Granted, this is

TABLE 13.1

Matrix of Relevant SJT Item Characteristics and Cultural Dimensions

<i>Hofstede's (1991) Cultural Dimensions</i>	<i>Items Situations (Item Stems)</i>	<i>Response Options</i>	<i>Response-Option Effectiveness (Scoring Key)</i>	<i>Response- Option- Construct Relationship</i>
Individualism/collectivism				
Masculinity/femininity				
Power distance				
Uncertainty avoidance				

important because both value orientations theory and attribution theory show that the effectiveness and thus the scoring key of for SJT response alternatives will differ across cultures. Yet, SJT design also entails gathering job-related situations and possible response options to these situations. It would be interesting to investigate whether SMEs in different countries provide the same critical situations and response alternatives. This can be easily done by presenting a given job to SMEs in different countries and asking them to generate critical incidents. Such research can shed light on whether the same types of situations occur across cultures. As noted, we expect that cultural inhibitions are often so strong that specific situations and responses in one culture would never occur in another. Apart from scrutinizing the relevance of problem-situations and response options across cultures, the frequency of response-option endorsement should also be compared across cultures. Finally, it should be examined whether the relationship linking SJT items and SJT intended constructs is transportable across cultures. It might be important to investigate the cultural transportability of the item–construct relationship because it suggests that the extent to which an SJT can successfully be used across cultures is dependent on the nature of the test content vis-à-vis the similarities and differences between cultures with respect to that content. If the aforementioned is correct, then within one SJT, some SJT responses will be more cross-culturally valid than others depending on the (dis)similarity between cultures in item content.

Second, very little is known about which SJT features increase or reduce cultural differences. McDaniel and colleagues (chap. 9, this volume) provides a good review of various item characteristics that might impact on the criterion-related validity of SJTs in a national context. Yet, virtually none of these characteristics have been investigated in an international context. One exception is the presentation format of SJT items. Chan and Schmitt (1997) showed that a video-based presentation format significantly reduced Black–White subgroup differences as compared with a written format. Therefore, similar to research on cognitive-ability tests (Cole, 1981; Scheuneman & Gerritz, 1990), future studies should identify specific types of item characteristics that may moderate differential item functioning across cultures.

Third, future studies should go beyond examining the effects of culture on response-option choice and include the effects on criterion-related validity. As already noted, we retrieved only one study (Such & Schmidt, 2004) that examined the criterion-related validity of SJTs in a variety of countries. In that specific study, the SJT development and scoring followed an imposed etic approach as the SJT was developed in one country and then used in other countries, with the result being that the SJT was predictive

13. INTERNATIONAL SJTs**293**

in only half of the countries. Probably, there was a lack of overlap between the SJT and the criterion in the other half of the countries. In this chapter, we posited that future studies should use an emic approach so that the SJT scoring key is tailored to the specific countries where the criterion data are gathered, guaranteeing sufficient overlap between predictor and criterion domains. Future research should be conducted to test these ideas.

In a similar vein, there is a clear need for studies that examine how the criterion-related validity of SJTs might be influenced by differences across cultures in how the various performance dimensions are weighted and combined into an overall job-performance rating. For instance, if managers in a particular culture (country) value that people get along (a contextual performance dimension), the fact that an SJT in another culture (country) predicts well individual task performance does not say much about the relevance of this SJT for hiring the best personnel in that specific culture (country). To our knowledge, no studies have investigated these issues.

Fourth, there is a need for a priori theory-driven examinations of the impact of cultural differences on SJT performance and criterion-related validity. So far, previous investigations (i.e., Nishii et al., 2001) have correlated SJT responses that had already been gathered across various countries with country scores on Hofstede's (1991) dimensions. Apart from its a posteriori nature, another limitation of this approach is that individual and country levels of analysis are confounded because an individual in a given country is equated with the score of his or her country. Clearly, such a country score on a cultural dimension such as individualistic serves at best as only a proxy of an individual's standing on this cultural dimension. A better approach would consist of determining a priori which items might be prone to cultural differences. In addition, respondents' individual scores on Hofstede's (1991) scales or similar scales (e.g., the GLOBE project; House et al., 2004) should be gathered. Whitney and Schmitt (1997) published an excellent example of such an a priori theory-driven approach for examining the influence of culture on selection procedures. On the basis of prior theory about value differences between Blacks and Whites, they a priori determined biodata items that would be vulnerable to cultural differences between Blacks and Whites. Next, they measured the cultural values of the individual respondents and correlated them with the individuals' response selection on these items. Some support was found for the hypothesis that cultural values were associated with the observed difference in Black-White response choices.

Fifth, no studies have used a construct-driven approach for examining the cross-cultural validity of SJTs. As already mentioned, it would be

particularly interesting to examine whether *g*-loaded SJTs (i.e., SJTs whose content is loaded with cognitive ability) are more likely to exhibit cross cultural validity, all other things being equal, than SJTs that are less *g*-loaded. Future studies can test this proposition at various levels. In particular, researchers might test this proposition at the overall score level (e.g., by comparing a “cognitive” SJT with a “noncognitive” SJT) and/or at the item level (e.g., by comparing “cognitive” SJT items with “noncognitive” SJT items).

Finally, it should be noted that virtually all “international” SJT applications discussed in this chapter were conducted in the United States and/or in western Europe. Future studies should be conducted in other parts of the world. Only in that case, we can obtain a full understanding of the cultural influences on SJTs.

IMPLICATIONS FOR PRACTICE

A first practical lesson to be learned is that general statements such as “SJTs are useful [or not useful] in other cultures” are not warranted. Instead, the specific application should be taken into account. We showed that SJTs carefully developed in Singapore, Belgium, or Germany for predicting job performance in these countries might have validities in the same range as SJTs developed in the United States. So, practitioners should be aware of the type of international application of SJTs. If the SJT is used for within-culture applications (predictor and criterion data are gathered in the same culture, e.g., an organization in Korea hires Korean individuals for a given job in Korea, see Quadrants A and D in Fig. 13.1), cultural differences do not seem to be a major threat. The reverse is true for cross-cultural applications of SJTs (predictor and criterion data are gathered in different cultures, e.g., a multinational hires individuals for a given job in a host culture, see Quadrants B and C). In these applications, the cultural transportability of SJT item characteristics might be at risk. Hence, their criterion-related validity might suffer if practitioners do not ensure predictor and criterion overlap.

How might practitioners ensure predictor and criterion overlap in cross-cultural applications of SJTs? Generally, there are two solutions possible. One solution might consist of changing the criterion. For example, one might consider evaluating the expatriates by corporate personnel in the original culture. However, this solution is both practically and conceptually debatable. From a conceptual point of view, this would mean that the criterion is changed on the basis of the predictor. One should always take the primacy of the criterion into account. From a practical point of view, it

13. INTERNATIONAL SJTs**295**

does not seem very acceptable that corporate headquarters determine what is good and bad performance in a specific host country. Instead, inspection of expatriate success criteria indicate that good performance implies that the expatriate is evaluated positively in the host culture in terms of task and interpersonal performance domains.

Another solution might be to change the predictor (the SJT) and to tailor the SJT to each specific culture (country). This means that both item stems and response alternatives should be carefully scrutinized for clarity and relevance in the host culture. In addition, it does not make sense to use or develop a “universal” scoring key as the same response option might be effective in one culture and ineffective in another culture. Instead, organizations should invest time and money to determine the effectiveness of the various response options in different cultures. Accordingly, it should be possible to tailor the scoring key to the specific host culture so that the key is consistent with the specific cultural norms.

On a more general level, our recommendations to scrutinize the content of item stems and response alternatives and to develop culture-specific scoring keys question the utility of an imposed etic approach when developing SJTs. Instead, our recommendations are in line with an emic approach. Ascalon, Schleicher, and Born (2004) provided an example of such a tailored country-specific approach. They developed an SJT for selecting expatriates targeted to five countries (The Netherlands, China, Germany, the United States, and Spain). Their SJT consisted of written scenarios representing the interaction of the five nationalities with one another. The SJT was designed to measure empathy and ethnocentrism; two dimensions that were posited to be related to cross-cultural social intelligence. People from these five countries served as experts to determine how the response options scored on these two dimensions.

EPILOGUE

Recently, organizations have started to use SJTs in an international context. The use of contextualized measures such as SJTs in an international context puts some challenges for organizations on the table. This chapter posited that three factors might determine the cross-cultural validity of SJTs, namely the transportability of the SJT items characteristics, the matching of predictor and criterion domains, and the type of constructs measured.

One of the key premises was that using an SJT in a different culture than originally intended is conceptually not different from using an SJT for another job or occupation than originally intended. This meant that

the generalizability of SJTs to other contexts might be jeopardized if these measures were used in a different context (e.g., job, organization, culture) and for a different criterion than originally intended. This leads to two implications. First, the interpretation of the correct or appropriate behavioral response to a specific situation might differ as a function of cultural values. In other words, the scoring key might differ from one culture to another. Second, SJTs might have differential validity across cultures if SJT scores do not match the criterion data gathered in another culture. In cross-cultural applications of SJTs, tailoring the scoring key to the host culture might be a way of matching predictors and criteria.

ACKNOWLEDGMENT

I would like to thank David Chan for his valuable suggestions on an earlier version of this chapter.

REFERENCES

- Adler, N. J., Doktor, R., & Redding, S. G. (1986). From the Atlantic to the Pacific century: Cross-cultural management reviewed. *Journal of Management*, 12, 295–318.
- Ansbacher, H. L. (1941). German military psychology. *Psychological Bulletin*, 38, 370–392.
- Arthur, W. Jr., & Bennett, W. Jr. (1997). A comparative test of alternative models of international assignee job performance. In D. M. Saunders (Series Ed.) & Z. Aycan (Vol. Ed.), *New approaches to employee management*, Vol. 4: *Expatriate management: Theory and research* (pp. 141–172). Stanford, CT: JAI Press.
- Ascalon, M. E., Schleicher, D. J., & Born, M. P. (2004). *Cross-cultural social intelligence: The development of a theoretically-based measure*. Manuscript in preparation.
- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1–26.
- Behrmann, M. (2004). *Entwicklung und Validierung eines Situational Judgment Tests für Call Center Agents—CALCIUM25: Baustein für die Personalauswahl*. [Development and validation of an Situational Judgment Test for call center agents]. Unpublished dissertation, Universität Mannheim, Germany.
- Berry, J. (1969). On cross-cultural comparability. *International Journal of Psychology*, 4, 119–128.
- Bhawuk, D. P. S., & Brislin, R. W. (2000). Cross-cultural training: A review. *Applied Psychology: An International Review*, 49, 162–191.
- Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, 74, 478–494.
- Bond, M. H. (1983). A proposal for cross-cultural studies of attribution processes (pp. 157–170). In M. H. Hewstone (Ed.), *Attribution theory: Social and applied extensions*. Oxford: Basil Blackwell.
- Born, M. P. (1994). Development of a situation-response inventory for managerial selection. *International Journal of Selection and Assessment*, 2, 45–52.

13. INTERNATIONAL SJTs**297**

- Born, M. P., Van der Maesen de Sombreff, P., & Van der Zee, K. I. (2001, April). *A multimedia situational judgment test for the measurement of social intelligence*. Paper presented at the 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.
- 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.
- Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410–417.
- Cole, N. (1981). Bias in testing. *American Psychologist, 36*, 1067–1077.
- Cropanzano, R. (1998, April). *Organizational justice and culture*. Paper presented at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment, 6*, 115–123.
- Furrer, O., Liu, B. S., & Sudharshan, D. (2000). The relationship between culture and service quality perceptions: Basis for international market segmentation and resource allocation. *Journal of Service Research, 2*, 355–371.
- Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially available aptitude tests. *Journal of Applied Psychology, 77*, 298–308.
- Highhouse, S. (2002). Assessing the candidate as a whole: An historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55*, 363–396.
- House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Thousand Oaks, CA: Sage.
- Jin, Y., & Wan, Z. (2004, August). *Managerial competence oriented situational judgement tests and construct validation*. Paper presented at the 28th International Congress of Psychology, Beijing, China.
- Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology, 86*, 984–996.
- Kleinmann, M., & Strauss, B. (1998) Validity and application of computer-simulated scenarios in personnel assessment. *International Journal of Selection and Assessment, 6*, 97–106.
- Hofstede, G. (1991). *Culture and Organizations: Software of the mind*. London: McGraw-Hill.
- Kluckhohn, F. R., & Strodtbeck, F. L. (1961). *Variations in value orientations*. Evanston, IL: Row, Peterson.
- Lee, S., Choi, K. S., & Choe, I. S. (2004, July). *Two issues in situational judgment tests*. Paper presented at the annual convention of the American Psychological Association, Honolulu, HI.
- Lievens, F. (in press). Personnel selection research in an international context. In M. M. Harris (Ed.). *Handbook of research in international human resource management*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*, 442–452.

- Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, 10, 245–257.
- 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., 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.
- McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 9, 103–113.
- Morris, M. W., & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *Attitudes and Social Cognition*, 67, 949–971.
- Motowidlo, S. J., Diesch, A. C., & Jackson, H. L. (2003, April). *Using the situational judgment test format to measure personality characteristics*. Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.
- Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in Industrial Psychology* (pp. 241–260). Palo Alto, CA: Davies-Black Publishing.
- Mount, M. K., Witt, L. A., & Barrick, M. R. (2000). Incremental validity of empirically keyed biodata scales over GMA and the five factor personality constructs. *Personnel Psychology*, 53, 299–323.
- Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: multivariate framework for studying test validity. *Personnel Psychology*, 50, 823–854.
- Newell, S., & Tansley, C. (2001) International uses of selection methods. In C.L. Cooper & I.T. Robertson (eds.) *International Review of Industrial and Organizational Psychology*, vol 21. pp. 195-213. Chichester, Wiley: UK.
- Nishii, L. H., Ployhart, R. E., Sacco, J. M., Wiechmann, D., & Rogg, K. L. (2001, April). *The influence of culture on situational judgment test responses*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.
- Ployhart, R. E., & Ryan, A. M. (2000, April). *A construct-oriented approach for developing situational judgment tests in a service context*. Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.
- Ployhart, R. E., Wiechmann, D., Schmitt, N., Sacco, J. M., & Rogg, K. L. (2003). The cross-cultural equivalence of job performance ratings. *Human Performance*, 16, 49–79.
- Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counter-productive performance to global ratings of job performance: A policy capturing approach. *Journal of Applied Psychology*, 87, 66–80.
- Ryan, A. M., McFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology*, 52, 359–391.
- Salgado, J. F. (1997). The Five-Factor model of personality and job performance in the European Community. *Journal of Applied Psychology*, 82, 30–43.
- Salgado, J. F., & Anderson, N. R. (2002). Cognitive and GMA testing in the European Community: Issues and evidence. *Human Performance*, 15, 75–96.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology*, 56, 573–605.

13. INTERNATIONAL SJTs

299

- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., De Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology*, 88, 1068–1081.
- Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66, 310–328.
- Scherer, K. R., Wallbott, H. G., & Summerfield, A. B. (Eds.). (1986). *Experiencing emotion: A cross-cultural study*. Cambridge: Cambridge University Press.
- Scheuneman, J., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, 27, 109–131.
- Schmidt, F. L., & Hunter, J. E. (1984). A within setting test of the situational specificity hypothesis in personnel selection. *Personnel Psychology*, 37, 317–326.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.
- Schmidt, S. M., & Yeh, R. S. (1992). The structure of leader influence: A cross-national comparison. *Journal of Cross-Cultural Psychology*, 23, 251–264.
- Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach*. Thousands Oaks, CA: Sage.
- Schuler, H., Diemand, A. & Moser, K. (1993). Film scenes: development and construct validity of a new aptitude assessment method [In German]. Filmszenen: Entwicklung und Konstruktvalidierung eines neuen eignungsdiagnostischen Verfahrens. *Zeitschrift für Arbeits- und Organisationspsychologie*, 37, 3–9.
- Schwartz, S. H., & Bardi, A. (2001). Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross Cultural Psychology*, 32, 268–290.
- Schwartz, S. H., & Sagiv, L. (1995). Identifying culture-specifics in the content and structure of values. *Journal of Cross-Cultural Psychology*, 26, 92–116.
- Shackleton, V., & Newell, S. (1997). International assessment and selection. In N. Anderson & P. Herriot (Eds.), *International handbook of selection and assessment*. New York: Wiley.
- Smith, P. B., Dugan, S., Peterson, M. F., & Leung, K. (1998). Individualism/collectivism and the handling of disagreement: A 23 country study. *International Journal of Intercultural Relations*, 22, 351–368.
- Smith, P. B., Peterson, M. F., Schwartz, S. H., Ahmad, A. H., Akande, D., Andersen, J. A., Ayestaran, S., Bochner, S., Callan, V., Davila, C., Ekelund, B., Francis, P-H., Graversen, G., Harb, C., Jesuino, J., Kantas, A., Karamushka, L., Koopman, P., Leung, K., Kruzela, P., Malvezzi, S., Mogaji, A., Mortazavi, S., Munene, J., Parry, K., Punnet, B. J., Radford, M., Ropo, A., Saiz, J., Savage, G., Setiadi, B., Sorenson, R., Szabo, E., Teparakul, P., Tirmizi, A., Tsvetanova, S., Viedge, C., Wall, C., & Yanchuk, V. (2002). Cultural values, sources of guidance, and their relevance to managerial behavior: A 47-nation study. *Journal of Cross-Cultural Psychology*, 33, 188–208.
- Steiner, D. D., & Gilliland, S. W. (2001). Procedural justice in personnel selection: International and cross-cultural perspectives. *International Journal of Selection and Assessment*, 9, 124–137.
- Such, M. J., & Schmidt, D. B. (2004, April). *Examining the effectiveness of empirical keying: A cross-cultural perspective*. Paper presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.
- Tett, R. P., Jackson, D. N., & Rothstein, M. G. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44, 703–742.
- Trompenaars, F., & Hampden-Turner, C. (1997). *Riding the waves of culture: Understanding cultural diversity in business*. London: Nicholas Brealey.

**Au: F & S not
instead Pls.
check**

- Van Leest, P. F., & Meltzer, P. H. (1995). *Videotesting of social, leadership, and commercial competencies*. Paper presented at the 7th European Congress on Work and Organizational Psychology, Győr, Hungary.
- Whitney, D. J., & Schmitt, N. (1997). Relationship between culture and responses to biodata employment items. *Journal of Applied Psychology*, 82, 113–129.