Repositório ISCTE-IUL

Teaching Psychological Measurement: Taking into Account Cross-Cultural Comparability and

Cultural Sensitivity

Christin-Melanie Vauclair, Diana Boer, and Katja Hanke

Ask your students to imagine that they are taking an intelligence test and are asked the

following questions[1]:

1. If BAD is written 214, how would you write DIG in the same secret writing? _____

2. What does it mean if someone says "she's buffed"?

    a. She's got a cute rear-end; b. She's overweight; c. She's wearing leather; d. She's got

    polished manners.

3. Which word is most out of place here?

a. splib; b. blood; c. gray; d. spook; e. black.

4. What number comes next in the sequence, one, two, three, _____?

5. We eat food and we _____ water.

If your students come from a Western culture and have a middle-class background, they

probably found it easy to answer question 1. However, they may have struggled to figure out the

meaning of the words in question 2 and 3, which are taken from intelligence tests designed for

African Americans in the United States (Dove, 1971; Redden & Simon, 1986). These items

assessed the knowledge of African American "street language" and students familiar with this

subculture usually scored well on these tests. Yet average white middle-class college students

have been sometimes scored so poorly that they could be classified as intellectually disabled

_____

[1] Correct answers: 1. 497; 2. b; 3. c.; 4. many; 5. eat.

according to "street norms." Your students may also be surprised to learn that their answers to questions 4 and 5, which appear to be very simple, are wrong in this test. These two questions come from an intelligence test developed for the Edward River Australian Aboriginal community in North Queensland.[2] The correct answer to question 4 is "many," because in *kuuk thaayorre* language, counting only goes to three: *thana*, *kuthir*, *pinalam*, *mong*, etc. The word *mong* is best translated as "many," because it can mean any number between 4 and 9. The correct answer to question 5 is "eat," because there is no distinction between "eating" and "drinking" in the *kuuk thaayorre* language – the same verb is used to describe both functions. How many of these questions did your students get right?

This little exercise provides a good starting point for class discussions on psychological measurements and culture. You could ask students how they would feel if these kinds of tests were used as standardized intelligence tests; or if such a test would be used to decide whether they should be admitted to a graduate school or hired for a new job. Such a discussion stimulates students´ perspective taking and makes them aware of the cultural knowledge they usually take for granted. They may not be sensitive to the biases that are part of standardized intelligence tests, because their own middle-class background does not disadvantage them for these tests. By taking these culture-specific intelligence tests, which make non-mainstream cultural assumptions, students can come to experience some of the difficulties and issues involved with culturally biased methods of testing intelligence.

The issue of cultural bias is not restricted only to intelligence tests. In fact, psychological tests play an integral part in Western societies, helping decision-makers make informed decisions in many different domains (e.g., development of new policies or psychological diagnoses and treatments). Furthermore, measuring psychological characteristics (e.g., attitudes) is at the heart

---

[2] http://www.wilderdom.com/personality/intelligenceCulturalBias

of most quantitative psychological studies. It is crucial that assessment of these characteristics is reliable (consistent across time, individuals, and researchers) and valid (i.e., measuring what it is intended to measure). The development of reliable and valid psychological measurements is the fruit of a rigorous research enterprise in which the measures are subjected to various tests. The objective is to ascertain that something is measured well enough to have scientific validity for the population in which the test is applied. This process becomes somewhat more complex when the population is culturally diverse and the aim is to develop or use a psychological measure that is not culturally biased.

In this chapter we will emphasize the importance of culture when it comes to psychological measurements and identify some key measurement concepts in this context. Taking into account culture in psychological measurements brings with it a whole host of methodological issues that go beyond monocultural studies. Due to space constraints we expect that students are already familiar with basic measurement theory (classical test theory) and its key concepts (reliability and validity). In the end, we also provide examples of teaching activities that can be used to explain cultural concepts in psychological measurement.

The Role of Culture in Psychological Measurement

The 21st century is an era of increased cultural diversity due to globalization and greater facility of travelling, living, and working elsewhere. Given the fact that societies have become more multicultural it becomes crucial to develop psychological measurements that are culturally inclusive. Another important aspect is the fact that an increasing amount of *research* is today directed at understanding the role of culture in influencing different aspects of people´s behavior, thought, and attitude. In fact, one of the main quests of cultural and cross-cultural psychology is to gather evidence from different cultures to better understand which aspects of the human mind

and behavior are universal or culture-specific. However, in order to draw scientifically valid conclusions about cultural differences, the measurement of people´s mind and behavior must be accurate, which is not as straightforward as it is for monocultural studies.

It may be helpful to remind students that measurement is actually pervasive in our everyday lives. We measure our weight by stepping on a bathroom scale, and judge whether we have a fever by using a thermometer. However, when psychologists undertake measurements, they are usually interested in assessing mental capacities and processes which are called *latent psychological constructs* (e.g., intelligence, personality, attitudes) because they are somewhat hidden and elusive. This renders their measurement a tricky undertaking because one cannot use scales or thermometers to directly assess these constructs. Instead psychologists use a systematic procedure for assigning scores to individuals so that these scores represent the characteristic of interest. This procedure, called *psychometrics*, has developed in Western cultures and it bears some challenges when individuals come from another culture. In fact, it has become clear that research results and psychological measures are not always valid in Non-Western cultures.

The validity of research results for populations other than those that have been studied is referred to as *external validity*. In other words, it is the degree to which research results can be generalized beyond one´s sample. This is a crucial concept when dealing with different cultures. Psychologists often assume that certain aspects of the human mind and behavior are universal; however, the vast majority of psychological studies have relied on samples from so-called WEIRD societies (Western, Educated, Industrialized, Rich, and Democratic; Henrich, Heine, & Norenzayan, 2010), which challenges the assumption that the results are generalizable. In fact, Henrich and colleagues showed that some of the key findings in psychology, which were based on WEIRD samples and assumed to be universal, do not generalize to samples from other

cultures. Nowadays, an increasing number of studies compares multiple cultural groups with each other on psychological variables of interest in order to examine cultural similarities and differences (Matsumoto & van de Vijver, 2011).

Emic and Etic Approaches

In the psychological study of culture, two strategies can be distinguished. The first strategy looks at one culture specifically and investigates its specific characteristics, expressions, behaviors, or ways of thinking. In-depth bottom-up analysis reveals this culture's psychological constructs, which can be used for describing the culture, to develop interventions or to derive measures. This approach is called *emic* and it intends to reveal culture-specific constructs; it does not intend to uncover universals or culturally comparable constructs. The researchers discover the structure and meaning of concepts instead of predicting or imposing them. For example, culture-bound syndromes are culture-specific psychological disorders which can only be fully understood within a specific cultural context. This is the case for the culture-specific syndrome of *hikikomori* in Japan (Sakai, Ishikawa, Takizawa, Sato, & Sakano, 2004). It describes mainly male adolescents or adults who completely withdraw from social life, often seeking extreme degrees of isolation and confinement over a period of several months or even years. *Hikikomori* resembles different DSM-IV-R syndromes in Western cultures, such as social phobia, obsessive-compulsive disorder, depression and schizophrenia, but does not fit into a single category and has several culture-specific manifestations. This example underscores the importance of conducting culture-specific analyses of psychological phenomena.

One specific emic approach investigates indigenous cultures and is called Indigenous psychology. Yang (2000), a renowned Chinese indigenous psychologist, argued that current mainstream psychology is dominated by a Western perspective, so that Western psychological

theories are culture-specific and not applicable to other parts of the world. In fact, current mainstream psychology could be seen as a Western Indigenous psychology. Yang further suggested building a cross-cultural indigenous research agenda in order to develop a more balanced global psychology.

The second strategy, called *etic*, aims to uncover psychological universals. This approach is mostly quantitative, views culture as an independent variable, and does not intend to include culture-specific features. The researchers determine the structure of psychological constructs, and the main aim is to compare cultures in order to test theories or hypotheses. Hence, etic studies attempt to establish the external validity of the phenomena studied. Both approaches come with advantages and disadvantages, and this is why Berry (1989) proposed to combine their strengths in order to derive psychological constructs that are applicable in a multitude of cultures. He called the approach *derived etics*, which first discovers emic constructs in various cultures and then combines their comparable features into a derived etic construct. This approach contrasts with another possible strategy in cross-cultural psychology, which is called *imposed etics*. It uses constructs derived in only one culture (generally Western) which are then assumed to be applicable in other cultures.

Cheung, van de Vijver and Leong (2011) in their study of personality argued for a combined emic-etic approach, which (similarly to Berry's derived etic) is an integrative strategy aiming to describe a psychological construct covering universal as well as culture-specific aspects. The combined emic-etic approach has been used, for example, to establish the South African Personality Inventory (SAPI). The SAPI project showed that the Big Five factor structure of personality, developed in the U.S., was well represented in South Africa, but culture-

specific social and relational aspects of personality were also revealed (Cheung, van de Vijver, & Leong, 2011).

One may ask now which approach to take when conducting (cross-)cultural psychological research. The decision on whether to use an emic, etic, or a combined emic-etic approach will largely depend on the topic of research: whereas some psychological concepts are more universal, others may be truly culture-specific. Furthermore, the specific research question may imply that one approach is more appropriate than another. Emic approaches advance our understanding of psychological mechanisms rooted in one particular culture, whereas culture-comparative etic approaches assume similarities in psychological phenomena across cultures and can at the same time provide insight into cultural differences regarding the expression of these phenomena based on cultural values, environmental factors, or other culture-relevant components (e.g., crisis, history, political development).

Issues of Comparability – The Role of Bias and Equivalence

One of the most common ways of comparing cultures is to employ surveys in the form of anonymous questionnaires, where participants are asked for their responses on rating scales to a series of questions (Heine, 2008). The great challenge for self-report measures lies in using psychological measures that are *equivalent* and unbiased across cultures, so that scores or associations between variables can be compared across cultures. Measurements may not be comparable across cultures, for example, because they have not been translated well, the items do not measure the same latent construct across cultures, culture-specific concepts do not exist in other cultures (e.g., the concept of counting beyond three), or individuals from some cultures show a specific response style (e.g., tending to use the midpoint of the response scale). We will come back to these issues in more detail below.

Test bias is a form of systematic (i.e., non-random) error which leads to the phenomenon that respondents from one culture have an unwarranted advantage over respondents from another culture. There are various sources of test bias in cross-cultural research (construct, instrument, and method bias) which can jeopardize the cross-cultural validity of a measure. If a measure is biased, it threatens the equivalence of measurement outcomes across cultures, meaning that the results may not be comparable across the sampled cultures and are, therefore, uninterpretable, or even misleading. In fact, bias and equivalence are two sides of the same coin: cross-cultural equivalence requires the absence of bias and the presence of cross-cultural bias results in some form of non-equivalence (van de Vijver & Leung, 1997). In order to make sure that we are not comparing apples and oranges when comparing measurement outcomes across cultures, we must ascertained that the measures are valid, not biased, and therefore equivalent across cultures.

The intelligence test example above illustrates how individuals from certain subcultures will probably score better on this test than the average white middle-class college student, because the test assesses cultural knowledge that is not accessible to the latter. As such it is neither fair nor scientifically valid to conclude that white middle-class students´ low scores on this test mean that they are unintelligent. In this case, the *external validity* of the test is compromised by its lack of cross-cultural equivalence as well as the presence of test bias.

A Taxonomy of Biases

Construct bias. *Construct bias* means that the construct measured is not identical across cultures. It refers to the cultural specificity of a psychological construct and means that there is an incomplete overlap of definitions of the construct across cultures, differential appropriateness of (sub)test contents (e.g., specific skills do not belong to the repertoire of one of the cultural groups), or inadequate sampling of relevant contents and incomplete coverage of the construct

(van de Vijver, 2013). The latter is also referred to as *domain under-representation*, which occurs when a measure misses important aspects of a construct in a specific cultural setting. For example, intelligence is defined as including social competence in some Asian cultures, such as Taiwan, Japan and China, and therefore should be incorporated in tests assessing intelligence in these cultures (Smith, Fischer, Vignoles, & Bond, 2013). This is also an issue of *content validity*, which refers to the extent to which a measure presents all facets of a given construct. One way to overcome construct bias is *cultural decentering* -- a procedure through which cultural specific contents are removed and items are formulated independently from the context, so that the instrument's appropriateness is maximized for all involved cultural groups (van de Vijver, 2013).

Construct bias can compromise the *construct validity* of a measure in a specific cultural population. Construct validity refers to the degree to which a variable has been operationally defined so that it captures the essence of a latent psychological construct. For example, is the intelligence test above a good measure of intelligence across different cultures? If the answer is yes, the measure has good construct validity. If the answer is no, any cross-cultural research conclusions drawn from the use of the measure are limited (Woold & Hulsizer, 2011).

Method bias. Even if a construct is identical across cultures, method bias may be an issue. Method bias includes the administration process, the instrument itself, and the sampling. These have been discussed as the main sources of method bias in cross-cultural research, while response sets are still disputed in the literature as a form of method bias or a true cultural phenomenon (i.e., a culture-specific communication style).

Administration bias. Administration bias includes influences due to the administration of the given instrument (e.g., the test environment, instructions, tester/interviewer effects, communication problems). Even if test situations are kept constant across participants, bias may

occur because of uncontrollable events. For example, administration bias can arise when data collection takes place at the respondents´ home and is disrupted by noise or other interfering events. Another source of administration bias can occur due to the test instructions. It is possible that differences in the instructions (e.g., one cultural sample may need more explanation in comparison to another) can lead to an overall method bias because the test conditions are not held constant across the cultural samples.

Administration bias is directly related to the issue of *internal validity* of the research results. Internal validity in cross-cultural research refers to the extent to which it is possible to draw the conclusion that culture has a causal effect on the observed phenomenon. In fact, a primary concern in psychological studies is to control all extraneous or confounding variables that may be part of an administration bias. However, achieving internal validity is difficult in cross-cultural research because these studies are quasi-experimental by nature and, therefore limited in regard to cause and effect conclusions in the first place. Culture cannot be manipulated by the researcher because participants are already enculturated when they participate in the study. Moreover, using standardized administration procedures in some cultures can even introduce a cultural bias, if the test situation is not compatible with local customs and cultural standards. For example, indigenous psychologists in the Philippines have pointed out that standardized test administrations limit the interaction between participants and researchers, which is an undesirable test situation for Filipino respondents (Pe-Pua, 2006). They prefer a casual and non-directed conversation that is driven by the respondent rather than the researcher. The methodology that corresponds to this cultural preference has been called *pagtatanong-tanong* which means "casually asking around." Hence, using a standardized procedure may jeopardize the validity of the research results in this culture.

Sampling bias. Sampling bias means that samples are not comparable and variations in samples from one cultural context to the other can confound the observed scores. For example, if educational levels in minority and majority members of a sample are not controlled or corrected for, a comparison of psychological constructs will confound cultural and educational differences (van de Vijver, 2013). Since realistically speaking random sampling rarely occurs, researchers have to be careful when interpreting their findings and making attempts to generalize.

Instrument bias. This kind of bias relates to instrument characteristics (van de Vijver, 2013). An instrument developed in a Western setting then exported to a Non-Western context may cause issues with item familiarity, response modes (e.g., familiarity with computers), or response formats (e.g., multiple-choice formats; He & van de Vijver, 2013).

Response styles. Response styles (also referred to as response sets) are systematic tendencies to respond to questions in a particular way, mostly to give a good impression of oneself. This is especially the case when participants are giving answers to self-report measures, such as attitudes. Four different kinds of response styles have been studied. A very common impression management strategy to portray oneself in a favorable light is referred to as social desirability, which has been widely studied (Paulhus, 1991). Additionally, there are other styles of impression management, such as the tendency to agree regardless of the content of the questions (acquiescence response style), the tendency to use the extreme end points of a scale (extreme response style), and the tendency to overuse the middle point of a scale (midpoint response style).

The challenge of how to manage different response sets has been a subject for debate for decades. Usually response sets are considered a part of measurement error (Johnson, Kulesa, Cho & Shavitt, 2005). Thus, some have suggested correcting for response sets in survey research (He

& van de Vijver, 2013, 2015), because they can have an impact on the measurement structure, means of scales, and associations between variables (Welkenhuysen-Gybels, Billiet, & Cambré 2003). Whereas the more conventional perspective views response styles as a nuisance factor that should be corrected (see Hui & Triandis, 1989), others hold the view that response styles are a reflection of culture-moderated communication filters (Smith, 2004) and response sets thus have meaning and do not need to be corrected for (He & van de Vijver, 2013). For example, research has shown that East-Asian cultures favor mid-point response options (Van de Gaer et al., 2012) and that cultural values can explain some of the cultural response tendencies (Johnson et al., 2005). Consequently, the cultural context adds another layer of complexity to response sets.

Recent research questioned the debate about whether cultural response styles are nuisances which need to be corrected. In large-scale studies, He and van de Vijver (2013, 2015) provided empirical evidence that correcting for response sets does not necessarily increase the validity of cross-cultural comparisons and concluded that response sets are often used as communication styles in order to moderate or amplify responses.

Item bias. Item bias refers to poor item translation, inadequate item formulation (e.g., complex wording), nuisance factors (e.g., item(s) may invoke additional traits or abilities), incidental differences in appropriateness of the item content (e.g., topic of item of educational test not in the curriculum in one cultural group), and cultural specifics (e.g., connotative meaning and/or appropriateness of item content). For example, van de Vijver (2013) argued that if two people from different cultures have the same levels on the latent construct (e.g., they have the same level of intelligence), but their responses result in differences in mean scores on the measures, then it is very likely that the items are biased.

In sum, a number of different biases can hamper the meaning and validity of cross-cultural research results. In order to overcome some of these biases an adequate cultural adaptation and translation of the measure is crucial.

Translation and adaptation

Adequate translation is key to ensuring the validity of the measurements used in different cultural settings. The most commonly used method to translate a source survey into a target survey is the translation-back-translation procedure and translations are carried out as part of an "Ask-the-Same-Question" (ASQ) model (Harkness, 2003). However, this procedure is not recommended anymore, because it has some serious limitations (see Mohler, Dorer, de Jong, & Hu, 2016). Harkness used an example of how translation-back-translation procedures can fail: the German item "Das Leben in vollen Zügen genießen" can be translated by a naïve translator into "Enjoy life in full trains!" This translation is literal and has face-value. However, because this is a German expression of enjoying life to the fullest, the literal translation would simply be wrong.

The procedure for a translation-back-translation approach consists of various steps. First, the initial translation is carried out for a specific target population. This translation is then back-translated into the source language (e.g., source language: English; target language: Chinese). The next step is to compare the two translations of the source language in order to find any translation issues. The cross-cultural survey guidelines recommend producing the best possible translation and then making a direct evaluation in the target language[3]. According to the guidelines, translation-back-translation is considered an indirect comparison which is vulnerable to misleading insights and eventually to lower quality translations (see Harkness, et al., 2010).

---

[3] http://ccsg.isr.umich.edu/index.php/chapters/translation-chapter

A recommended alternative to the translation-back-translation procedure is the team approach, also referred to as the *committee approach*. It requires a team of knowledgeable bilingual experts who produce the best possible translation in a team effort and directly evaluate the solution within the committee. The translation of the measure is completed when the committee achieves a consensus about the appropriateness of the translation.

The cultural adaptation of measures goes beyond mere translation (Behr & Shishido, 2016). The issues of appropriateness of research material needs to be considered here specifically, because it will affect the quality of cross-cultural research. For example, imagine you are asking your students to respond to the following question: "To which religious group do you belong?" and you provide them with a categorical response option regarding the religious groups *Muslim Sunni, Muslim Shia, Hinduism, Bahá'i Faith, Sikhism, Shintoism, Daoism, Traditional*, *none* and *other (specify)*. Assuming that your students have a Western background, you can discuss with your students the appropriateness of this question in their cultural setting.

A key question in the cultural adaptation of psychological measures is whether it captures the psychological construct adequately, representatively, and comprehensively. Hence, an important consideration in this process is to reflect carefully from which cultural context the measure is coming (source) and to which culture it should be applied (target). For example, a survey that originated from the U.S. and includes questions about schooling and politics that only make sense in this setting becomes meaningless when used with another cultural sample. Hence, it is important to adapt the items in such a way that they make sense in the target population. According to Leung and van de Vijver (1997) and van de Vijver and Leung (2011), there are two ways that this can be achieved: (1) by adapting the material in a way that it is adequate for the culture of interest or, (2) by assembling and designing a new instrument (i.e., referred to as

indigenization). The problem with the latter option is that it is difficult to make cross-cultural comparisons if the instrument is different.

Equivalence. Why can't we simply compare measurement scores between cultures and interpret the difference as culturally meaningful? The reason is that we first need to be very sure that the same underlying latent construct was measured and hence is indeed comparable. If the construct is meaningless, differently defined, or not comparable (i.e., inequivalent), comparisons are baseless. Equivalence is defined as the level of comparability of scores across cultures. When testing the equivalence of psychological measures across cultures, there is a hierarchy of different levels of equivalence that can be statistically examined: functional, structural, metric, and scalar equivalence.

Functional equivalence requires an in-depth understanding of each cultural context and extensive qualitative and conceptual work. It is the most abstract level of equivalence and, therefore, difficult to proof whether the exact same constructs are captured exhaustively. This type of equivalence also taps into issues of *linguistic equivalence*, which refers to the extent to which the construct of interest has been adequately translated. Functional equivalence is often assumed and not tested, whereas the remaining levels heavily rely on being statistically tested and should not be assumed (van de Vijver & Poortinga, 1997).

Structural equivalence – sometimes also referred to as construct equivalence - refers to indicators that tap adequately the construct of interest in a culturally meaningful way. Those indicators need to be relevant and representative of the construct in each cultural setting. This level of equivalence is the basis for all cross-cultural comparisons and needs to be established before higher levels of equivalence can be tested (see Table 1).

Metric equivalence is needed in order to compare relative patterns in the data (e.g. correlation-based analyses) between two or more cultural groups. Psychometric tests are employed to identify and remove problematic items from the scale. Metric equivalence is defined as having the same measurement unit across cultures, but not the same origin (He & Van de Vijver, 2013). This means thatit is possible to compare correlations and investigate associations with regressions, but it is not possible to make mean-level comparisons.

*Scalar (*or *full score) equivalence* is the highest level of equivalence and the most difficult one to establish statistically. It occurs when measures have the same measurement unit and origin. Only if scalar equivalence is established, is it possible to directly compare mean scores between two or more cultural groups using t-tests or analysis of variance.

Multigroup confirmatory factor analysis is the most common procedure to test for structural, metric and scalar equivalence. More recently, other procedures have been suggested that seem to be more appropriate for testing equivalence across cultures (e.g., Bayesian structural equation modelling; for an overview, see Boer, Hanke, & He, in press).

In this chapter, we have discussed the role of culture in psychological research and what we need to consider if we want to make meaningful comparisons across cultures. We introduced the different approaches to studying culture in psychology: emic and etic approaches and the fact that derived etic and combined emic-etic approaches seem most fruitful, because they embrace universality and culture-specificity at the same time. We have summarized the different kind of biases that can occur and what can be done to reduce bias. We have also summarized the controversial role of response sets in cross-cultural research. Since adequate translation is one important way to ensure item quality and reduce item bias, we explained why the committee approach is the most recommended procedure. Furthermore, we introduced the different levels of

equivalence (functional, structural, metric and scalar equivalence) which should be tested before any cross-cultural comparisons are made. In the following section, we suggest some useful teaching exercises to apply the contents of this chapter.

Exercises and Examples

Exercise 1: Translation and Linguistic Equivalence

Let the students get together in teams of 2 or 3 and ask them to translate some items using translation-back-translation and the committee approach. Discuss and compare the translations using the two different approaches. This assignment requires bilinguals.

Exercise 2: Distinguishing Culture-Specific (Emic) and Universal Concepts (Etic)

This exercise aims to sensitize students for culture-specific and universal concepts with cross-cultural differences.

Task 1. Ask the students to identify a music style specific to their local area or region. The students then describe the origin and characteristics of the music as well as the specific context in which people listen to it. Although most music styles are hybrids with many different cultural influences, they are usually not directly comparable to other music styles due to their historical and regional specificity. Note that there are regions in the world where music is not described in terms of *genres* as is the case in China. Here, bands (particularly Chinese-Pop or, for short, C-Pop bands) are phenomena that are similar to genres. In these contexts, bands rather than particular music genres could be described.

Task 2. The next exercise is suitable for multi-cultural classrooms. Students are asked to identify a characteristic of people in their country, which seems distinctive to their cultural identity. The group then discusses whether this characteristic is indeed a culture-specific construct. There is a chance that most mentioned characteristics overlap across cultures. A

discussion may then reveal that the actual culture-specific aspect lies in the importance that is attributed to cultural characteristics in specific contexts. This means that the construct can be universal, whereas its significance and manifestation in specific contexts can vary across cultures. One example would be that the Brazilian national identity seems particularly linked to Samba and to music in general. Brazilians are not the only culture in which music is associated with national identity and in which Samba is liked. However, the *strong meaning* of Samba is culture-specific in Brazil.

Exercise 3: Developing a Culturally Sensitive and Widely Applicable Measure

This exercise puts the contents of this chapter into practice. The task is to develop a research plan for development of a psychological measure that captures the functions of music listening. The measure should be culturally sensitive as well as applicable in a wide range of different cultures. Groups of 3-6 students should work on this task. The task given to students may read as follows: *Your task is to develop a culturally sensitive measure that captures "the functions of music listening". The measurement is supposed to be applicable to young people around the globe. How would you come up with the items? Who would you ask to provide input? Which functions and which contexts of music listening would you consider? The research agenda may entail different stages and research methodologies.*

One possible research agenda solving this task is presented in the following example. This example can be used for discussing and evaluating the proposals provided by student groups. A critical reflection of the provided example could also be part of the in-class discussion.

Example: Process of Developing a New Culturally Sensitive and Widely Applicable Measure

Aiming to derive a holistic framework of functions of music listening, the second author used a mixed-methods design (Boer, 2009; see also Boer & Fischer, 2012; Boer et al., 2011, 2012). The research agenda followed a two-stage procedure. First, a qualitative culturally decentred study set out to identify the reasons why people like to listen to music, the meaning of music in their personal and social life as well as in their families and cultures. The data were used for development of a psychological measure which was then validated in a second stage.

The first research stage utilized a qualitative online survey aiming to capture various personal, social and cultural functions of music from people hailing from different cultures and nations. This strategy intended to maximize the cultural diversity in regard to music usage and rituals, while trying to avoid domain-underrepresentation of certain music functions. Multiple questions were used to capture each of three contexts of music listening functions:

- personal: *What does music mean to you? Please write your thoughts about the role music plays in your life. How does music influence your life? Think about one specific situation when you were listening to music in the last three days. Please describe what you thought, felt, and did in that situation.*

- social: *What role does music play when you are hanging out with your friends? What is the meaning of music for your family members?*

- cultural: *What is the meaning of music in your home country? What is the meaning of music in your cultural community?*

Each participant answered three of those questions (one on each context), which were presented as open-ended questions without space limitations. The questionnaire was available in English and German.

The qualitative study was conducted online with samples of young people from four cultural clusters: South America, (South) East Asia, Anglo-Saxon regions, and Europe. The link to the survey was disseminated via discussion boards of music-related websites, because people highly committed to music were sought to provide most valuable insights about their involvement with music and the reflected reasons for music listening in various contexts. In total, 222 participants from 31 countries contributed their qualitative answers to this study. The data were analyzed using thematic analysis. The analysis revealed seven main functions of music listening: music in the background, memories through music, music as diversion, emotions and self-regulation through music, music as reflection of self, and social bonding through music (for more detail see Boer, 2009; Boer et al., 2011). For each main function, multiple sub-functions were identified.

In the second research stage, the aim was to develop a measure that captures these seven main functions of music. The qualitative input of study 1 was used to generate items. Item generation was mostly based on phrases provided by the participants of the qualitative study. An initial set of 229 items was derived. Their usefulness and applicability was assessed in a multicultural committee, who checked all items for clarity and translatability as well as congruence and consistency in content. The remaining 74 items were translated from English to German and Spanish using the translation-back-translation approach.

The set of 74 items was administered in quantitative surveys in Spanish, English and German to student participants in Mexico, New Zealand and Germany (see Boer, 2009; Boer et al., 2011). The parallel assessment of the measure in three languages and cultural groups aimed to enhance the cultural sensitivity and cross-cultural applicability of the resulting item selection. The data were analyzed using Principal Component Analysis. Items with clear factor loadings

were identified and items were removed that had double loadings, formed single factors or that loaded inconsistently across the three samples. This resulted in 36 final items measuring 10 functions of music listening: Emotions, Social Bond with Friends, Family Bond, Venting, Dancing, Background, Focus, Values, Political Attitudes, and Cultural Identity. In order to assess the cross-cultural comparability of the measure´s factor structure, its equivalence was tested across the sampled cultures. The results showed that the scale meets structural equivalence across the three cultural samples (see Boer, 2009; Boer et al., 2012). The scale was named RESPECT-Music[4] (Ratings of Experienced Social, PErsonal, Cultural Themes of Music Functions).

The learning goals of this exercise are threefold. First, students will discuss in-depth which issues need to be considered when developing and testing a new culturally sensitive measure. The familiarity with the topic (music listening) will provide a fairly easy entry point for in-depth discussions based on personal experiences. Second, this task engages students in considering different cultural perspectives in the development of one measure that aims to be applicable across cultures. Third, the complexity of developing culturally sensitive scales will become apparent. The class can discuss and evaluate the different approaches and solutions developed by the groups as well as the exemplary solution presented here. This will emphasize the fact that a research question may be answered in different ways.

---

[4] The RESPECT-Music scale in its current form is available in 9 languages so far and has been assessed in 15 countries. Download available at https://www.uni-koblenz-landau.de/de/koblenz/fb1/institut-psychologie/abteilungen/sozial%20und%20organisationspsychologie/research/

References

Behr, D., & Shishido, K. (2016). The Translation of Measurement Instruments for Cross-cultural Surveys. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (pp. 269–287). London, UK: Sage.

Berry, J. W. (1989). Imposed etics-emics-derived etics: The operationalization of a compelling idea. *International Journal of Psychology, 24*, 721-735. doi:10.1080/00207598908246808;

Boer, D. (2009). *Music makes the people come together. Social functions of music listening for young people across cultures*. (Doctoral dissertation, Victoria University of Wellington, New Zealand). Retrieved from: http://researcharchive.vuw.ac.nz/bitstream/handle/10063/1155/

Boer, D., & Fischer, R. (2012). Towards a holistic picture of functions of music: A culturally decentred qualitative approach. *Psychology of Music*, *40*, 179-200. doi:10.1177/0305735610381885

Boer, D., Fischer, R., de Garay Hernández, J., González Atilo, M. L., Moreno, L., Roth, M. & Zenger, M. (2011). The functions of music-listening across cultures: The development of a scale measuring personal, social and cultural functions of music. In F. Deutsch, M. Boehnke, U. Kühnen, & K. Boehnke (Eds.), *Rendering borders obsolete: Cross-cultural and cultural psychology as an interdisciplinary, multi-method endeavor*. Bremen, Germany: International Association for Cross-Cultural Psychology.

Boer, D., Fischer, R., Tekman, H. G., Abubakar, A. A., Njenga, J., & Zenger, M. (2012). Young people's topography of musical functions: Personal, social and cultural experiences with music across genders and six cultures. *International Journal of Psychology*, *47*, 355-369. doi:10.1080/00207594.2012.656128

Cheung, F. M., van de Vijver, F. J., & Leong, F. T. (2011). Toward a new approach to the study of personality in culture. *American Psychologist*, *66*(7), 593. doi:10.1037/a0022389

Dove, A. (1971). The "Chitling" Test. In L. R. Aiken, Jr., *Psychological and Educational Testings*. Boston: Allyn and Bacon.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010). Designing questionnaires for multipopulation research. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B-E., Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multicultural, and multiregional contexts* (pp. 33-58). Hoboken, NJ: John Wiley and Sons.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35-56). Hoboken, NJ: John Wiley & Sons.

Heine, S. J. (2008). *Cultural Psychology*. New York: W.W. Norton.

He, J., & Van de Vijver, F. J. R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G. A. D. Liem & A. B. I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney* (pp. 39-56). Charlotte, NC: Information Age Publishing.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61-83-135. doi:10.1017/S0140525X0999152X

Redden, P. M., & Simons, J. A. (1986). *Manual for the Redden Simons "Rap" Test*. Ankeny, IO: Des Moines Area Community College.

Matsumoto, D. & Van de Vijver, F. (Eds) (2011). *Cross-cultural Research Methods in Psychology*. Cambridge, UK: Cambridge University Press.

Mohler, P., B. Dorer, J. de Jong, & M. Hu. (2016). *Translation: Overview. Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Pe-Pua, R. (2006). From decolonising psychology to the development of a cross-indigenous perspective. In U. Kim, K. S. Yang & K. K. Hwang (Eds), *Indigenous and cultural psychology: Understanding people in context* (pp. 109-137). New York: Springer.

Sakai M., Ishikawa S., Takizawa M., Sato H., Sakano Y. (2004). The state of hikikomori from a family's point of view: Statistical survey and the role of psychological intervention. *Japanese Journal of Counseling Science*, 37, 168–179.

Smith, Peter B., Fischer, R., Vignoles, V. L. & Bond, M. H. (Eds.). (2013). *Understanding Social Psychology Across Cultures: Engaging With Others In A Changing World* (2nd ed.). Thousand Oaks, California: Sage.

Van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology, 43*, 1205–1228. doi: 10.1177/0022022111428083

Van de Vijver, F. J. R. (2013). Construct bias. In K. Keith (Ed.), The encyclopedia of cross-cultural psychology (pp. 233-235). Hoboken, NJ: Wiley. doi:10.1002/9781118339893

Van de Vijver, F. J. R. (2013). Method bias. In K. Keith (Ed.), *The encyclopedia of cross-cultural psychology* (pp. 878-880). Hoboken, NJ: Wiley.

Van de Vijver, F. J. R. (2013). Item bias. In K. Keith (Ed.), The encyclopedia of cross-cultural psychology (pp. 772-774). Hoboken, NJ: Wiley.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.

Van de Vijver, F. & Leung, K. (2010). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. Van de Vijver (Eds), *Cross-cultural Research Methods in Psychology* (pp. 17-45). Cambridge, UK: Cambridge University Press.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29-37. doi:10.1027/1015-5759.13.1.29

Woolf, L. M. & Hulsizer, M. R. (2011). Why diversity matters: The power of inclusion in research methods. In K. D. Keith (Ed.), *Cross-cultural psychology: Contemporary themes and perspectives* (pp. 56-72). Oxford, UK: Wiley-Blackwell.

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*, 241–264. doi:10.1111/1467-839X.00067

Table 1. *Overview of the different types of equivalence, their sources of bias and analytical procedures (adapted from Boer, Hanke, & He, in press)*

Type of Equivalence	Definition	Source of Bias	How to establish this type of equivalence?
Functional Equivalence	The same construct exists in the cultural groups	Construct has different functions or is interpreted differently across cultures	Appropriate translation and domain representation
Structural Equivalence	Same indicators can be used to measure a construct in different cultural groups	Rough factor loading structure is not similar across cultures – some factor loadings are not significant in some cultural groups and in others it is	Analytical tools For example: Exploratory Factor Analysis, Multigroup Confirmatory Factor Analysis
Metric Equivalence	Indirect comparisons possible between groups, such as correlation-based analysis (e.g. regression models), but no multilevel models	Factor loadings are not identical across cultures	Some analytical tools: Strict (exact) test: Multigroup Confirmatory Factor Analysis (same factor loadings) Approximation test (tolerant towards trivial measurement unit differences): Bayesian Structural Equation Modeling
Full score equivalence	Direct comparisons of (latent) mean scores across cultures are possible (e.g. t-tests, ANOVA, etc)	Factor loadings and indicator intercepts differ across cultures	Some analytical tools: Strict (exact) test: Multigroup Confirmatory Factor Analysis (same intercepts) Approximation test (tolerant towards trivial measurement unit differences): Bayesian Structural Equation Modeling