International Perspectives in Psychology: Research, Practice, Consultation
2016, Vol. 5, No. 1, 51–65

© 2015 American Psychological Association
2157-3883/16/$12.00 http://dx.doi.org/10.1037/ipp0000042

Adaptation of Assessment Scales in Cross-National Research: Issues, Guidelines, and Caveats

Barbara M. Byrne
University of Ottawa

Increasingly, over the past 2 decades, there has been a growing interest in cross-national comparisons. This activity, in turn, has precipitated an escalating number of assessment scales being translated into other languages for use in countries and cultures that differ from those of the original scales (typically developed and normed in the United States). Recent criticism of these translated scales has highlighted the singularity of focus on linguistic equivalence albeit with little to no regard for equivalence of the measured constructs, relevance of item content, familiarity with item format, and insufficient rigor of the methodological strategy, thereby leading to serious biasing effects that ultimately yield a multiplicity of complexities in cross-national research and practice. Intended as an aid to researchers confronted with the task of translating and adapting an assessment scale for use in a country and culture that differs from that of the original scale, this article (a) highlights the critical importance of equivalence as it relates to the translated and adapted scale, in addition to the construct(s) it is designed to measure, (b) identifies the major threats to such equivalence and exemplifies several ways by which they can bias cross-national comparisons, (c) outlines a recommended series of psychometric analytic stages that can lead to both a close translation and a rigorously adapted assessment scale, (d) describes and explicates the hierarchical set of steps necessary in testing equivalence of the adapted instrument within and across national groups, and (e) presents the advantages and disadvantages of the adaptation approach recommended for use in this article.

Keywords: cross-cultural comparisons, cross-national comparisons, multigroup measurement/structural equivalence, translated/adapted assessment scales

Over at least the past two decades, there has been a rapidly growing interest in cross-national comparisons. Although such comparisons traditionally have been the province of cross-cultural psychologists, a review of the extant literature reveals this investigative work now to be of substantial interest within mainstream psychology with researchers, whose work has previously been conducted within a single culture, now extending across national borders (van de Vijver & Hambleton, 1996). This phenomenon mimics the "transport and test" approach noted by Berry (1969) and extends far beyond the boundaries of cross-cultural psychology. Over and above three journals singularly devoted to the discipline of cross-cultural psychology (*Journal of Cross-cultural Psychology*, *Journal of Cross-cultural Management*, and *International Journal of Intercultural Relations*), van de Vijver, Chasiotis, and Breugelmans (2011) note that all mainstream psychology journals have published articles reporting on cross-cultural studies and that many other journals publish such articles on a regular basis. One critically important outgrowth of this work has been the burgeoning number of psychological assessment scales being translated into multiple languages. Indeed, based on statistical data from PsycINFO (F. J. R. van de Vijver, personal communication, April 22, 2015), van de Vijver (2009) reported the number of journal articles addressing translation/adaptation issues during this 20-year period to have increased at least 350%!

This article was published Online First October 19, 2015.

I am grateful to Kurt Geisinger for his invaluable suggestions on an earlier draft of this article. I also thank Merry Bullock, Bonnie Nastasi, and members of the 2012 and 2013 APA Committee on International Relations in Psychology (CIRP) for their support and insightful comments during its embryonic stage of development.

Correspondence concerning this article should be addressed to Barbara M. Byrne, School of Psychology, University of Ottawa, Ottawa, ON K1N 6N5, Canada. E-mail: bmbch@uottawa.ca

Based on the earlier work of van de Vijver and Leung (1997), He and van de Vijver (2012) note that there are basically three options from which to choose in development of a measuring instrument for use in cross-national research: *adoption*, *adaptation*, and *assembly*. Of the three, adoption represents the simplest and least expensive option as the task focuses solely on linguistic equivalence between the translated (i.e., target) scale and the original (i.e., source) scale. The major disadvantage of this option, however, is the validity of its use if, and only if, it is known that there is a perfect alignment between both the measuring instrument and the construct it is designed to measure in the target country (a questionable feat unless specifically tested and found to be so). The second option, adaptation, is much broader in scope than adoption. In addition to the need for close translation, it also includes all psychometric activities that should take place in practice when developing an assessment scale that has been constructed in one language and culture for use in a different language and culture (Hambleton, 2005). One excellent example of this option is the work of Aegisdóttir and Einarsdóttir (2012) in their adaptation of the Beliefs about Psychological Services scale for use in Iceland (I-BAPS). The third option of assembly actually represents the building of a new linguistically and culturally appropriate scale. Presented with failure of both the adoption and adaptation options in yielding a scale capable of demonstrating acceptable linguistic, cultural, and psychometric accuracy, this option remains the only choice (He & van de Vijver). One very exceptional example of the assembly option is evident in the development of the South African Personality Inventory (SAPI; Valchev et al., 2013), an indigenous personality scale that not only addresses all 11 linguistic languages of South Africa, but also takes into account the existing cultural diversity of the country subsequent to Apartheid.

Provided with evidence of the rapidly expanding number of cross-national studies being conducted by mainstream psychologists and the contention that most of these researchers have not been trained in the cross-cultural tradition (van de Vijver et al., 2011), it seems likely that the primary purpose of these studies is the conduct of cross-national comparisons, albeit necessarily accompanied by the need to translate and adapt an existing assessment scale into another language for use in the target country and culture of interest. Of the three possible scale development options cited earlier, adaptation would appear to be the most probable choice under such circumstances. This option is therefore the one incorporated into the present article.

As a prerequisite to addressing the topic of assessment scale adaptation here, I consider it important to clarify and delineate my use of the phrase “cross-national,” which could be interpreted as a conflation of the terms “country” and “culture.” Whereas *cross-national research* can be defined as investigations conducted across two or more national political and language boundaries, *cross-cultural research* can be defined as investigations conducted within and across two or more cultural groups (see, e.g., Harkness, van de Vijver, & Mohler, 2003). In other words, all cross-national research can be classified as cross-cultural, but not all cross-cultural research can be classified as cross-national (e.g., Hispanic Americans compared with Asian Americans).

One common weakness in multigroup comparison research in general and cross-national research in particular is the pervading assumption that both the assessment scale and the construct(s) it is designed to measure are operating equivalently across the groups of interest. (A more detailed elaboration of this issue is addressed later.) When research is based on use of an assessment scale that has been solely translated, regardless of sound linguistic correspondence, there is substantial evidence that such nonequivalence is not uncommon even within national borders (see, e.g., Byrne, 1993; Guo, Suarez-Morales, Schwartz, & Szapocznik, 2009; van Leest, 1997). Thus, it should not be surprising to find even stronger evidence when the comparison groups are cross-national (see, e.g., Byrne & Campbell, 1999; Byrne & Watkins, 2003). Unquestionably, it cannot be assumed that an assessment scale developed for use in one language and country will automatically measure the same phenomena in exactly the same way when translated into another language for use in a different country and culture. There may be language differences that change the meaning of certain items, cultural differences in targeted behaviors believed to represent an underlying construct, experiential differences in respondents' exposure to particular assessment scale formats, and the like. Any one of these features, in addition to many more equally troublesome dynamics, serves to underscore the fact that translating and adapting an assessment scale from one language and culture into another language for use in a different culture is much more complex than may appear on the surface.

The primary intent of this article is to assist researchers in avoiding the difficulties noted earlier by becoming knowledgeable in the development of a psychometrically sound adapted assessment scale, as well as cognizant of the many complexities commonly encountered in this process. More specifically, the purposes are fivefold: (a) to underscore the critical importance of equivalence as it relates to both the adapted measuring instrument and the construct(s) it is designed to measure in comparison of the national groups of interest; (b) to outline the major threats to such equivalence and exemplify several complexities that can impact the validity of cross-national comparisons; (c) to identify and describe the series of stages involved in the process of assessment scale adaptation; (d) to explain and describe the hierarchal set of steps necessary in testing for equivalence of the adapted instrument across national borders; and (e) to present the advantages and disadvantages of current assessment scale adaptation procedures as recommended by the International Test Commission (ITC, 2005). Although issues related to equivalence as well as translation/adaptation of assessment scales in the conduct of cross-national comparisons have been widely published, many have tended to focus on particular components of this process. In contrast, the present article is unique in presenting an integration of the full range of tasks required in establishing a psychometrically sound instrument for these purposes, including the steps involved in testing statistically for evidence of both measurement and structural equivalence.

The Issue of Equivalence in Cross-National Comparisons

In the conduct of cross-national comparisons, it is essential that both the construct and the related measuring instrument are operating equivalently across groups. As noted by He and van de Vijver (2012), without construct equivalence, there can be no cross-cultural comparisons. Construct equivalence can be evidenced through group consistency of its meaningfulness, as well as its structural dimensionality. Within the context of this article, the dimensional aspect of the construct is considered to represent structural equivalence.[1] An understanding of *structural nonequivalence* can be illustrated via a study by Byrne and Watkins (2003) in which they tested the equivalence of an assessment scale designed to measure physical (ability and appearance) and social (peers and parents) self-concepts for Australian and Nigerian adolescents. Evidence of structural nonequivalence was claimed on the basis of the inequality of construct relations (i.e., latent factor correlations) across the two groups. Specifically, the latent factor correlation between the Ability and Appearance dimensions of Physical Self-concept differed significantly, with the correlation being higher for Australian than for Nigerian adolescents. From a substantive perspective, this finding may derive from the differing societal values held by these two adolescent groups. Self-perceived physical attractiveness may be defined in terms of a superior body physique for Australian adolescents, albeit in terms of beautiful facial features for Nigerian adolescents.

Measurement equivalence, as the name implies, focuses on the assessment scale structure with concern being the extent to which it is operating equivalently across groups. Although there are several aspects of the scale that can be tested for multigroup equivalence, two are of primary concern and are the only ones mentioned here: (a) the regression link between each item and the latent factor construct to which it is assigned (i.e., the factor loadings), and (b) the item intercepts (i.e., the observed variable means). An example of *measurement nonequivalence* can be shown in a study by Byrne and Campbell (1999) designed to test for equivalence of the Beck Depression Inventory (BDI, 1999) across Canadian, Swedish, and Bulgarian adolescents. Results revealed 14 items to be

[1] As noted by van de Vijver (2011), there is some discrepancy between the quantitative and cross-cultural psychology literatures in use of the term "structural equivalence."

functioning differentially between the Canadian and European adolescents, albeit for only 4 items between the Swedish and Bulgarian adolescents. Although such nonequivalence reflects a difference in the perception of item content, it can also result from the impact of other complexities, a topic to which we turn in the next section.

To determine the extent to which a measuring instrument exhibits both structural and measurement equivalence across groups, its factorial structure is put to the test, statistically, based on a hierarchy of increasingly rigorous analytic steps. This procedure is presented in more detail following description of the assessment scale adaptation process.

Potential Complexities Impacting Cross-National Comparisons and the Issue of Bias

Testing for equivalence of the intended construct and its related assessment scale across national groups has shown that findings of nonequivalence are not uncommon. In fact, it is likely safe to say that such results are more often the case than not so. Most reasons for this common finding can be linked to complexities arising from test bias some of which, in turn, can evolve from the translation of an assessment scale into another language. Although it is possible that the final version of a translated scale may be found to be linguistically equivalent to its original source scale, this instrument can nonetheless sustain many critical limitations that appropriately argue against its use in cross-national research. These weaknesses stem from the failure to formally determine a priori: (a) similarity of the construct(s) being measured in terms of conceptualization, operationalization, dimensionality, and targeted behaviors; (b) familiarity with the scale item format in terms of required response (e.g., multiple choice, Likert scaling, dichotomous scaling); and (c) problematic item content with respect to interpretation and clarity of text (e.g., use of colloquialisms). Failure to investigate these potentially differential aspects between the translated and original instruments will lead almost certainly to the introduction of test bias and ultimately, to structural and measurement nonequivalence across the groups under study.

In psychometric terms, test bias conveys the notion that test scores based on the same items measure different traits and characteristics for each group. Viewed from a cross-cultural perspective, bias implies differences in the assessment scale that elicit differential meaning within, versus across national/cultural groups. As emphasized by van de Vijver and Tanzer (1997, p. 264), the issue of bias does not relate to the intrinsic properties of a measuring instrument per se, but rather, "to characteristics of a cross-cultural comparison of the instrument." As such, statements regarding bias always refer to use of an instrument within the framework of particular applications of cross-cultural comparison (van de Vijver & Leung, 2011). For example, whereas an instrument may reveal evidence of bias in a comparison of Canadians and Romanians, such evidence may not be present in a comparison of Canadians and Brazilians.

In general, problems of bias in cross-cultural research can be linked to three primary sources: (a) the construct of interest (construct bias), (b) the methodological procedures (method bias), and (c) the item content (item bias). (Readers wishing greater detail concerning test bias, in addition to suggestions for ways to address the problem, are referred to van de Vijver & Leung, 1997)

Construct Bias

This first type of bias conveys the notion that the construct being measured holds some degree of differential meaningfulness across the cultural groups under study thereby running counter to the assumption that the measured construct is equivalently relevant across the groups of interest. It seems likely that many of the discrepancies associated with construct bias derive from the process of enculturation. In a recent consolidation of previous work addressing this topic, van de Vijver and Leung (2011) conclude that construct bias can be linked to two primary sources: (a) the incomplete overlap of construct-relevant behaviors, and (b) the differential appropriateness of behaviors in different cultures. Of the two, the first source is the more pervasive as restricted behavior overlap can arise in similar, albeit slightly different ways. Given that example instances can typically portray conceptual notions better than word descriptions, illustrations of each are now presented.

Incomplete overlap of construct relevant behaviors. This first example draws from Ho's (1996) work on filial piety, the concept of being a "good" son or daughter. This construct has been the focus of much discussion in cross-cultural research concerned with comparisons between Western and non-Western societies, as the number of behaviors considered relevant to this construct is widely divergent. Although the Chinese, for instance, hold the expectation that children will take full care of and responsibility for their elderly parents, their perception of such behavior is much broader than is the case in Western societies. Relatedly then, an instrument designed to measure filial piety based on the Chinese perception of this construct would tap into aspects of filial piety that are unrelated to those pertinent to the Western notion. Likewise, an inventory developed, say, in the United States and based on the Western notion of the construct would be unlikely to cover all aspects of the construct as it relates to the Chinese notion. Thus, any comparisons made across these two national groups, based on either one of these instruments would be problematic.

A second example of incomplete construct overlap can be taken from the work of Cheung and colleagues (2001) in their attempt to replicate scores on the Five Factor Model (FFM) of personality based on several samples of Chinese respondents. Of relevance in this instance is a differential dimensional structure of personality across the cultural groups. Despite claims that the FFM represents a universal measure of personality, results based on Chinese samples consistently revealed six rather than five factors. These consistent findings led Cheung et al. (2001) to subsequently develop the Chinese Personality Assessment Inventory (CPAI), an indigenous instrument designed to take into account the sixth factor of "Interpersonal Relatedness" within the context of Chinese society. In an effort to verify this additional factor, these authors conducted a subsequent joint factor analysis of the CPAI and FFM; results substantiated their inability to subsume this sixth factor under the FFM (Cheung, 2012).

Differential appropriateness of construct relevant behaviors. For an example of how such disparity can occur, let's take the construct of practical intelligence, albeit in a more specific form that represents "street-smartness." Suppose that an inventory is designed and developed in Canada with the intent of being used to compare street-smartness of preadolescent schoolchildren across inner urban and rural Canadian communities. Let's further postulate that this instrument is subsequently translated for use in Pakistan. Indeed, it seems highly probable that the behaviors targeted as measurements of street smartness in Canada will differ substantially from those in Pakistan. Thus, although this instrument is used with the same age population and based on the same inner urban/rural comparisons, it would likely fail to capture common situations experienced by the Pakistani children and may yield less meaningful results.

Two studies provide additionally good examples of how behaviors associated with a construct can have a high probability of being differentially appropriate if tested across Western and non-Western cultural groups. Here again, the construct of interest is intelligence and each study identifies some aspect of social responsibility as an essential component of intelligence. The first example (Serpell, 2011), based on three decades of research conducted in Zambia, has shown consistent findings that in addition to cognitive alacrity, social responsibility (household/family responsibilities) constitutes a critical dimension of intelligence despite almost exclusive focus of the institutionalized school curricula on "the cultivation of knowledge and cognitive skills" (Serpell, p. 126). In the second study, Hein, Reich, and Grigorenko (2015) found community-oriented adaptive skills to be a substantially more important dimension of intelligence than academic skills for children living in Kenyan villages. Indeed, the findings reported in these two studies stand in stark contrast to Western nations wherein intelligence is typically school-oriented with tests of intelligence focused solely on scholastic ability.

Method Bias

This second type of bias is an umbrella term encompassing several aspects of the methodological strategy employed in the cross-cultural comparisons of interest. Although I present only three examples here, several others can be found in van de Vijver and Leung (2011). This first example focuses on *sample bias*, an incomparability of samples resulting from phenomena other than the target factors of interest. Consider a study designed to test for cultural differences in traditional family values (roles within the family; kinship ties) for Spaniards versus Brazilians based on a sample of university students in Madrid and Sao Paulo, respectively. Although ideally it would be preferable to randomly sample culturally representative respondents within each of these countries, this approach is very often not possible because of financial and/or other constraints. Hence, participants are recruited using convenience sampling, with gender and age customarily recorded for purposes of control. However, there remain other important background variables that can severely impact valid findings of cultural differences pertinent to Spain and Brazil. In both countries, the university is located in a large city and thus it seems likely that the sample will comprise three subcultural groups of students—long-term national residents, recent immigrant residents, and international short-term residents. Even with the exclusion of the international students, the Spanish and Brazilian samples are plainly mismatched. Although this disparity may be of less concern in the investigation of other constructs, it is of particularly serious concern here, given that family values are strongly rooted in their traditional cultural origin and, as a result, differ widely around the globe (see, e.g., Georgas, Berry, van de Vijver, Kaγitçibaşi, & Poortinga, 2006).

A second example of method bias can derive from problems associated with the assessment scale itself and is often termed *instrument bias*. More specifically, it relates to the differential response, by comparative groups, to the structured item format. One recognized source of bias here bears on the extent to which respondents are familiar with the item format. Given that many affective instruments are based on paper-pencil tests that are structured within the framework of a Likert scaling format, it is very possible that this type of stimulus response may be unfamiliar to some cultural groups thereby reflecting itself in a biasing of item scores. Still another type of instrument bias can be found with respect to patterns of response (i.e., response styles) and response sets (i.e., social desirability; acquiescence), particularly as they relate to personality and attitude scales (van de Vijver & Poortinga, 2005). For example, Hui and Triandis (1989) reported a strong tendency for Hispanics to choose the extreme categories of a 5-point Likert scale when compared with the Euro Americans, albeit such difference was not found when the scale was reconstructed as a 10-point scale.

A final example of method bias derives from the administration of an assessment scale and can be termed *administration bias*. This type of bias derives from some discrepancy associated with the administration of an assessment scale to the participants of comparative groups. For example, in a comparison of math self-concept for grade three students, let's suppose that all but one group of students were guided through the completion of a set of practice items related to a Likert-scaled self-report scale. The presence of practice items for some but not all respondents serves to make the testing process nonequivalent from the start. Although administration bias can derive from many sources and can distort all modes of testing, the interview format would appear to be particularly vulnerable.

Item Bias

This third and final category of bias refers to distortions at the item level. Items are said to be biased if they elicit differential meaning of their content across groups. Differential interpretation of item content by members of nationally and/or culturally different groups derives largely from a diversity of sociocultural contexts that include the family, the school, the peer group, and society at large. Let's take for example, an instrument designed to measure self-concept. Typically, items on such a scale require the respondent to evaluate him or herself in comparison with others who share the same social milieu (e.g., work, school, family).

Thus, it seems evident that diverse socialization practices cannot help but lead to different sets of criteria against which to judge one's perception of self. Within a cultural context, Markus and Kitayama (1991) posit that these criteria can be classified as representing *independent* versus *interdependent* perceptions of self, with the former focusing on the individual's separateness/independence and the latter on connectedness/interdependence to others. Whereas the independent self-construal is consistent with the perspective of Western cultures, the interdependent self-construal is consistent with that of non-Western cultures. Accordingly, characterization of the self in Western cultures is oriented toward a focus on the uniqueness of one's attributes thereby leading to claims of self-actualization, autonomy, and the like. In contrast, characterization of the self in non-Western cultures is defined in terms of one's social relationships, rather than the uniqueness of one's attributes. Of highest regard are one's efforts to "fit in, to belong . . . and in general, to become part of various social units" (Markus & Kitayama, 1991, p. 22), rather than to set oneself apart from the group. Given the dichotomy of these cultural frameworks, it seems likely that the basis of comparison by which American versus Japanese adolescents would judge themselves in response to particular items on a self-concept scale could differ dramatically. For example, presented with the following two Likert-scaled items, "I can do most things pretty well" and "I am assertive when I need to be,"[2] it seems probable that whereas the Americans would tend to select a high scale point consistent with their innate sense to affirm their autonomy, the Japanese would tend to select a more modest scale point yet without feeling any loss of autonomy in doing so.

Of the three types of test bias noted above, Method Bias is likely the simplest to conceptualize as its deficiencies are fairly straightforward and uncomplicated. In contrast, Construct Bias is far more complex and Item Bias slightly less so. Thus, with respect to the latter types of test bias, two crucial questions that need to be addressed are (a) how can construct and item biases be detected, and (b) how can they be averted?

Both types of test bias can arise as a consequence of weaknesses in the translation procedures used. At the same time, however, they can be detected and prevented using a comprehensive approach capable of identifying aspects of both the original and source languages and cultures that may trigger either construct and/or item bias. Although a detailed description and explanation of this information goes beyond the bounds of the present article, interested readers can find an excellent overview of the topic in van de Vijver and Leung (2011).

Over and above the issue of test bias, it is important to draw attention to one major methodological obstacle that can also severely impact the equivalence-testing process. This complexity arises mainly as a consequence of tests for equivalence that involve a large number of groups (e.g., typically $>$ 5). To date, research suggests that it is particularly relevant to cross-cultural data. In a study designed specifically to illustrate the extent to which use of the standard structural equation modeling (SEM) approach to testing for equivalence can be problematic when applied to large-scale cross-cultural data, Byrne and van de Vijver (2010) found it impossible to establish a well-fitting configural model (detailed later in this article), a necessary prerequisite to further equivalence testing of an assessment scale. Essentially, the core of this difficulty lies with the restrictiveness of the confirmatory factor analytic (CFA) model in requiring the nontarget loadings of the hypothesized model to be zero thereby addressing the assumption that the specified structure will fit equally well across all groups. Based on responses to the Family Values Scale (FVS; Georgas, 1989) for 5,483 university students from 27 countries, Byrne and van de Vijver (2010) found this assumption to be untenable given the extensive number and diversity of cross-group model modifications needed to attain a satisfactory level of model fit to the data. Indeed, results suggested that the more widely diversified the sociogeographical contexts of the groups, the more difficult the problem.

In an effort to address this procedural obstacle in the testing of assessment scale equivalence when the number of groups is large, Asparouhov and Muthén (2014) have recently advanced a new approach to fitting the multigroup configural model which they have termed the *alignment method*. The key to this new procedure lies in the relaxation of restrictions related to the nontarget CFA factor loadings noted earlier. In contrast to the standard strategy, the alignment method not only automates and greatly simplifies the equivalence testing process, but also provides a detailed account of parameter equivalence pertinent to every model parameter in every group (Asparouhov & Muthén, 2014).

[2] Taken from the Competence subscale of the Multidimensional Self Concept Scale (MSCS; Bracken, 1992).

The Process of Assessment Scale Adaptation

Unquestionably, transforming an assessment scale from one language and culture to another language for use in a different country and culture is definitely *not* a simple task. Rather, it is a very complex, tedious, and time-consuming endeavor that requires the work of many experts knowledgeable in the areas of linguistics and psychometrics, in addition to the specific substantive area pertinent to the constructs being measured. Although several different qualitative and quantitative strategies are possible in the translation of instruments, the process of adaptation, as it has evolved from the invaluable early work of Hambleton (1994, 2005) and colleagues (e.g., Hambleton & Kanjee, 1995; Hambleton & de Jong, 2003) is considered to represent the most comprehensive approach to transforming an instrument from its use in one cultural context to that in a different cultural context. Given that this adaptation approach encompasses all necessary activities involved in a psychometrically sound conversion process, it is now widely recognized as the gold standard in the evolution of establishing a culturally redesigned measuring instrument. Indeed, He and van de Vijver (2012, p. 4) avow that this option "has become so popular that adaptation has become the generic term to refer to the translation process of psychological instruments."

This adaptation approach derives from a concerted effort on the part of cross-cultural methodologists and psychometricians (e.g., Geisinger, 1994; Hambleton & de Jong, 2003; van de Vijver & Hambleton, 1996) to right the wrongs of the past 50 years in transforming instruments developed in one country (usually the United States) for use in another, based only on a translation from the source language to the target language. Guidelines detailing this adaptation process were initially developed by the ITC in the early 1990s (see Hambleton, 1994) and have subsequently been updated twice. The current version of the ITC *Guidelines on Adapting Tests* can be found at: http://www.intestcom.org. (For a comprehensive description, explanation, and historical review of these guidelines, together with a standardized review format in their use, see Hambleton, 2005; Hambleton & Zenisky, 2011.)

In a recent attempt to further improve this adaptation process, researchers (e.g., Cheung, 2012; Cheung, van de Vijver, & Leong, 2011) suggest that taking an emic-etic approach to this task can lead to a more appropriate and ultimately more valid measurement scale for use in the target country as it serves as a check not only on the appropriateness of the items, but also on the appropriate structure of the construct. Indeed, this result certainly rang true for adaptation of the Beliefs About Psychological Services scale for use in Iceland (IBAPS; see Aegisdóttir & Einarsdóttir, 2012). Unfortunately, however, this important adaptation approach may not always be possible because of ownership of the instrument by testing corporations and the related copyright restrictions. Thus, in the text that now follows, I describe the basic test adaptation process consistent with the most recent version of the ITC Guidelines. However, for a walk-through of an emic-etic approach to this process based on an actual application, I strongly recommend that readers consult the Aegisdóttir and Einarsdóttir (2012) article.

Translation Stage of Adaptation

Importantly, the adaptation approach uses backward translation as only one of several quality controls in the translation process. Appropriate implementation of this procedure demands the need for qualified translators who are fully competent in both languages of interest, are familiar with the cultures associated with each language group, have a sound grasp of the subject domain measured, and have a solid understanding of item and test construction (Hambleton & Kanjee, 1995). Taken together, an appropriate translation should represent "a balanced treatment of psychological, linguistic, and cultural considerations" (van de Vijver & Tanzer, 2004, p. 266). Consistent with this caveat, the ITC *Guidelines on Adapting Tests* focus on a three-step process: (a) the instrument is translated from the source to the target language, (b) the translated instrument is then translated back into the original language (back translation), and (c) using a team approach comprising independent teams of "qualified" translators, the three translated versions (i.e., original, target, back-translated) of the instrument are examined with a keen eye to their correspondence and resolution of any discrepancies that may be detected along the way (Hambleton, 2005). (For an example of the translation of a survey scale into three languages based on the ITC Adaptation Guidelines, readers are referred to Sireci, Yang, Harter, & Ehrlich, 2006.)

The intent of this initial step in the process of instrument adaptation is the establishment of a comprehensive and multifaceted translation of the instrument for use in the target country. Once this critical first step has been completed, the adaptation procedure continues and involves at least three additional steps: (a) pilot- and field-testing of the instrument, (b) validation of the adapted instrument's scores within the country in which it is to be used (the target country), as well as across the target and source countries, and (c) establishment of norms for use of the adapted instrument within the country of its intended use.[3] These three final steps are now briefly described.

Pilot and Field-Testing Stage of Adaptation

The intent of the pilot- and field-testing of an adapted instrument is to identify any poorly functioning items (Geisinger, 1994), thereby serving to pinpoint potentially biased items. Whereas pilot testing typically involves only a few trial administrations of the instrument to a small sample of individuals, field testing represents its administration to large samples that are representative of the target population. Both of these testing phases are invaluable in their identification of poorly constructed items. For example, the content may be grammatically incorrect, ambiguous, or include unintended colloquial expressions. Although each of these instances is problematic, the latter two are particularly critical as they serve as potential inhibitors of equivalence between the source and target versions of the assessment scale. Pilot and field-testing are also vital in their identification of items found not to be measuring their targeted subscale constructs as well as expected. Clearly, such difficulties also serve to undermine the possibility of equivalence. Each of these situations demands further examination and decision-making regarding either the modification or deletion of any identified malfunctioning items.

Given that differences in language and socialization contexts bear importantly on the interpretation and cognitive processing of item content, an important challenge in the adaptation of assessment scales is the extent to which these dynamics are consistent across respondents of different national/cultural groups. Evidence of such discrepancies is indicative of biased measurement (Harkness et al., 2003). In an effort to minimize these biasing effects, the practice of cognitive interviews has been included as a component of this early pretesting stage. Used either before or after pilot testing, this protocol represents a qualitative, semistructured procedure whereby a small sample of respondents is asked to explain their rationale in responding to particular test items. The overall purpose of cognitive interviewing is to ensure that the content of each scale item: (a) makes sense to all respondents, (b) is applicable to the life contexts of all respondents, and (c) has been interpreted and understood according to the meaning intended (Miller, Mont, Maitland, Altman, & Madans, 2011). Taken together, pilot testing, field-testing, and cognitive interviews all serve to minimize the item biasing effects. Of important note brought to my attention by a reviewer, many of the more recent large-scale international projects now use cognitive interviewing as an alternative to the traditional pilot testing procedures noted above.

Construct Validation Stage of Adaptation

Following completion of the pilot and field testing phase, the adapted instrument requires validation. Accordingly, the factorial structure of the instrument needs to be tested statistically in order to validate its hypothesized structure with respect to: (a) the number of underlying factors (representing the number of constructs), (b) the pattern of factor loadings (representing the items designed to measure each construct), and (c) relations among the factors (representing the extent to which the underlying constructs are correlated). These analyses address the issue of construct validity.

Confirmatory factor analysis (CFA) provides the most appropriate and statistically rigorous

[3] As noted by two reviewers, the norming of particular instruments is not always needed or appropriate. These may include measures that assess culturally relevant dichotomously displayed behaviors, as well as others that may be theory-driven (e.g., the Rorschach test).

method in testing for evidence of construct validity as it takes a hypothesized-testing (i.e., confirmatory) approach, rather than an exploratory (e.g., exploratory factor analysis) approach to analysis of the data. As such, the factorial structure of the instrument is postulated a priori and then tested for the validity of this hypothesized structure. All analyses are conducted within the framework of structural equation modeling (SEM) based on the related computer software. The process of establishing the construct validity of an adapted instrument logically involves two important stages: (a) testing within the country of intended use (the target country), and (b) testing across the new and former country where the instrument was originally developed (the source country). Although an in-depth description of the analyses involved in this construct validation stage goes beyond the bounds of this article, detailed explanation and description of CFA models, together with annotated illustrations of diverse applications based on different SEM programs, can be found in Byrne (2006, 2010, 2012a); for a brief description and comparison of the primary SEM software programs, see Byrne (2012b).

Construct validation within the target country. During the first construct validation phase, interest focuses on the factor structure of the adapted instrument based on data representing the country of its intended use. More specifically, the hypothesized factorial structure of the instrument is tested statistically with respect to the number of underlying factors, the pattern of factor loadings onto their targeted factors (the constructs) and intercorrelations among the construct dimensions (the factors), as noted earlier. In the event that there are discrepancies in the factor structure (e.g., an item may cross-load on a nontarget factor), the model under test is modified accordingly and then retested. This would also trigger a need to review this item's translated content.

Construct validation across target and source countries. Once construct validity of the adapted instrument has been established within the context of the target country, a logical next step in the validation process is to determine the extent to which the scale's factorial structure is equivalent to that of the original version for respondents in the source country. This process represents testing for the instrument's equivalence (i.e., invariance) across the target and source countries. This multigroup testing process, of course, is also pertinent for any desired future group comparison studies within the target country, as well as for other cross-national contexts. A brief overview of this equivalence testing procedure is now presented. These analyses are again conducted within the framework of SEM for two important reasons: (a) the systematic manner in which these tests for equivalence can be conducted and (b) because SEM enables a testing for structural as well as measurement equivalence. Indeed, van de Vijver (2011, pp. 27, 28) argues that "There is no other statistical theory that allows for such a fine-grained, flexible, and integrated analysis of equivalence." He further contends that "SEM has been instrumental in putting equivalence testing on the agenda of cross-cultural researchers and in stimulating the interest of cross-cultural studies."

Testing for Multigroup Equivalence

Testing for the equivalence (or invariance) of an assessment scale entails a hierarchical set of steps that should always begin with determination of a well-fitting *baseline* model for each group separately. Once these baseline models are established, their separate model specifications are combined thereby representing a multigroup baseline model. In technical terms, this initial multigroup model is termed the *configural model* (Horn, McArdle, & Mason, 1983) and is the first and least restrictive one to be tested. With the configural model, only the extent to which the same pattern (or configuration) of fixed and freely estimated parameters holds across groups is of interest and thus no equality constraints are imposed. It is this multigroup model for which sets of parameters are put to the test of equality in a logically ordered and increasingly restrictive fashion. In contrast to the configural model, all remaining tests for equivalence involve the specification of increasingly restrictive cross-group equality constraints for particular parameters.

Measurement equivalence is concerned with relations between the observed variables (i.e., the scale items), which are directly measurable, and their links to the unobserved (or latent) variables (i.e., the factors), which are not directly measurable. In testing for measurement equivalence, these parameters always include the factor loadings (regression of observed variables onto their related latent variables) and may include the observed variable error variances and any error covariances (commonly termed error correlations). Should a researcher be interested in subsequently testing for latent factor mean differences, then tests for measurement equivalence must include a test for the equality of the observed variable intercepts (i.e., item means) as such equality is assumed in tests for latent (i.e., factor) mean differences.

In contrast to measurement equivalence, structural equivalence (i.e., construct equivalence) focuses on the unobserved (i.e., latent) variables. As such, structural equivalence is concerned with the extent to which the meaning and dimensional structure of a psychological construct are identical across groups. These parameters always involve relations among the factors (i.e., factor covariances) and may extend to include the factor variances and error residual covariances. (For a more detailed explanation of these issues and procedures, as well as illustrated applications, readers are referred to Byrne, 2006, 2008, 2010, 2012a.)

Establishing the multigroup model. As a prerequisite to testing for multigroup equivalence, it is customary to determine a baseline model for each group separately. This model represents the one that best fits the data from the perspectives of both parsimony and substantive meaningfulness. Given that measuring instruments are often group-specific in the way they operate, baseline models may not be completely identical across groups (Byrne, Shavelson, & Muthén, 1989). For example, it may be that the best-fitting model for one group includes a cross-loading that may not be specified for the comparison group. Presented with such findings, Byrne et al. (1989) showed that by implementing a condition of *partial measurement invariance*, multigroup analyses can still continue given that the recommended conditions are met. As such, some but not all measurement parameters are constrained equal across groups in the testing for equivalence. A priori knowledge of such group differences is critical to the application of equivalence-testing procedures. Once a well-fitting baseline model has been established for each group separately, these final models are then combined in the same file to form the multigroup configural model.

Testing the configural model. The initial step in testing for cross-group equivalence requires only that the same number of factors and their loading pattern be the same across groups; as such, no equality constraints are imposed on the parameters. Of primary interest here is the extent to which the configural model fits the multigroup data. Importantly, despite evidence of good fit to the multisample data, the only information that we have at this point is that the factor structure is *similar*, but not necessarily *equivalent* across groups as equivalence of the factors and their related items have not yet been put to the test. That is to say only the overall model fit has been tested. Nonetheless, the configural model serves two important functions. *First*, it allows for equivalence tests to be conducted across the groups simultaneously. *Second*, the fit of this configural model provides the baseline value against which all subsequently specified models are compared in the test for cross-group equivalence.

Testing for measurement equivalence. When a researcher is concerned only with the extent to which the factorial validity of an instrument is equivalent across independent samples, measurement equivalence focuses solely on the invariant operation of the items and, in particular, on the factor loadings (commonly termed metric equivalence). As such, interest centers on the extent to which the content of each item is perceived and interpreted in exactly the same way across the samples. If a test for latent mean (i.e., factor mean) differences is of interest, then, it is imperative that the intercepts (i.e., the observed variable means) also be tested.

In testing for the equivalence of factor loadings, these parameters are freely estimated for the first Group only; the factor loading estimates for Group 2 (and any additional groups if this is the case) are constrained equal to those of Group 1.[4] Provided with evidence of equivalence, these factor loading parameters remain constrained equal across groups while simultaneously testing for the equivalence of additional parameters (e.g., item intercepts). On the other hand, confronted with evidence of nonequivalence related to particular factor loadings, equality constraints related to these parameters are released and tests for equivalence proceed if the data meet the recommended conditions of partial measurement equivalence (see Byrne et al., 1989).

[4] Determination of which group should serve as Group 1 is purely arbitrary.

In a comprehensive study of measurement equivalence, Meredith (1993) distinguished among three increasingly restrictive forms, which he termed weak, strong, and strict tests of measurement equivalence. Testing for the equivalence of factor loadings falls into the category of *weak* equivalence.[5] *Strong* equivalence occurs when both the factor loadings and intercepts are constrained equal across groups. Should a researcher wish also to include item error variances, tests for equivalence are then considered to represent *strict* equivalence.

Testing for structural equivalence. In contrast to tests for measurement equivalence, tests for structural equivalence focus on the unobserved (or latent) variables. As such, testing for multigroup equivalence of an assessment scale can involve both the factor variances and their covariances. However, the factor covariances are typically of most interest.

In summary, testing for assessment scale equivalence across groups is a hierarchically ordered process that begins with establishment of a multigroup model, termed a configural model, based on the baseline factorial structure established for each group separately. Given findings of a well-fitting configural model to the sample data, analyses proceed by testing first for measurement equivalence followed by tests for structural equivalence consistent with the imposition of equality constraints across groups as noted earlier. I conclude by presenting three circumstances in which the parameters of interest can vary. *First*, in testing for equivalence of the target and source scales across their national samples, primary interest would focus on the factor loadings and the factor covariances. In the event that latent factor mean differences were of interest, then the intercepts would also be included (see, e.g., Byrne, Stewart, Kennard, & Lee, 2007). *Second*, in testing for equivalence of the adapted instrument across various groups within the target country, as well as across other national groups, interest again would focus on the factor loadings and factor covariances and may include the item intercepts should latent mean differences be of interest. However, depending on the research issue being addressed, be it theoretical or empirical, there may be reason for additionally testing equivalence of the item error variances and possible error covariances, in addition to equivalence of the factor variances.

Norming Stage of Adaptation

The final phase in the adaptation process is to establish norms for the translated instrument in the country of its intended use (but see Footnote 2). Norms provide information about the placement of an individual score within a population of respondents to the same instrument. As such, it allows the test giver to interpret an individual test score and is important for diagnosis or placement (Geisinger, 1994). When an assessment scale developed in one country is adapted for use in another country, it may well be inappropriate for interpretation of scores to be based on normative information related to the source instrument as norms are relevant only to the population in which they are derived. This judgment depends on the nature of the original norm sample and the extent to which it was broadly inclusive of members of the target population. Thus, unless deemed inappropriate for the particular instrument involved (see Footnote 2), the adaptation of an instrument for use in another country should include the gathering of normative information relative to performance in the new country. Only then can the interpretation of scores have any valid meaning.

Advantages and Disadvantages of Adaptation

In general, there are at least three important advantages of the adaptation approach presented in this article compared with often-used separate translation and adaptation procedures. For an extended discussion of these advantages and disadvantages, readers are referred to Harkness et al., 2003). *First*, this approach addresses construct equivalence in addition to linguistic equivalence. *Second*, it uses a committee approach such that groups of specialists work both separately and together in determining the extent to which concepts, words, and expressions are "culturally, psychologically, and linguistically equivalent in the target language" (Hambleton, 2005, p. 4). *Third*, this approach utilizes linguistic experts who are fully bilingual in both the source and target languages, as

[5] There is some confusion regarding the assignment of Meredith's (1993) categorical labels within the cross-cultural literature where configural invariance is commonly referred to as "weak invariance" (see, e.g., Davidov, 2008).

well as being native citizens of either the target or source country.

Despite these three very constructive and progressive features of this adaptation approach, they are countered by two very major, yet realistic disadvantages. The first of these is a realization of the difficulty involved in acquiring well-qualified and knowledgeable experts in linguistics, psychometrics, and the relevant field of psychology. The second disadvantage is the obvious costs involved in the locating of experts capable of conducting such adaptation procedures and with the costs of implementing all stages of the adaptation process.

Conclusion

Multigroup comparisons across national borders can be fraught with problems, many of which are exceedingly subtle and not easily discerned. Of critical import in such inquiry is empirical evidence that the assessment scale and the underlying construct(s) it is designed to measure are operating equivalently across the groups of interest. Although such equivalence is often assumed in translated instruments considered to be linguistically equivalent to its original source, this assumption typically does not hold because of innumerable biasing effects that can impact both the construct and various aspects of the scale itself. Based on the comprehensive approach to adaptation outlined in this article many of these biasing effects can be identified and addressed. Through a description and exemplification of types of bias that can impede the quest for equivalence, a detailed overview of the adaptation approach to attainment of instrument equivalence initiated by Hambleton (1994) and now documented in the ITC Guidelines (www.intest.org), and a comprehensive outline of the equivalence testing procedure within the framework of SEM, I hope that I have succeeded not only in heightening awareness of the many complexities that accompany the conduct of multigroup comparisons across national borders, but also in assisting readers to better comprehend the manner by which these issues can be addressed.

References

Aegisdóttir, S., & Einarsdóttir, S. (2012). Cross-cultural adaptation of the Icelandic Beliefs About Psychological Services Scale (IBAPS). *International Perspectives in Psychology: Research, practice, Consultation, 1,* 236–251. http://dx.doi.org/10.1037/a0030854

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling, 21,* 495–508. http://dx.doi.org/10.1080/10705511.2014.919210

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4,* 199–228. http://dx.doi.org/10.1080/00207596908247261

Bracken, B. A. (1992). *Multidimensional Self Concept Scale (MSCS)*. Rolling Meadows, IL: Riverside Publishing.

Byrne, B. M. (1993). The Maslach Burnout Inventory: Testing for factorial validity and invariance across elementary, intermediate, and secondary teachers. *Journal of Occupational and Organizational Psychology, 66,* 197–212. http://dx.doi.org/10.1111/j.2044-8325.1993.tb00532.x

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Erlbaum.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20,* 872–882.

Byrne, B. M. (2010). *Structural equation modeling with Amos: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Erlbaum.

Byrne, B. M. (2012a). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Taylor & Francis/Routledge.

Byrne, B. M. (2012b). Choosing SEM computer software: Snapshots of LISREL, EQS, AMOS, and Mplus. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 307–324). New York, NY: Guilford Press.

Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology, 30,* 555–576. http://dx.doi.org/10.1177/0022022199030005001

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466. http://dx.doi.org/10.1037/0033-2909.105.3.456

Byrne, B. M., Stewart, S. M., Kennard, B. D., & Lee, P. (2007). The Beck Depression Inventory II: Testing for measurement equivalence and factor mean differences across Hong Kong and American Adolescents. *International Journal of Testing, 7,* 293–309. http://dx.doi.org/10.1080/15305050701438058

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

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34,* 155–175. http://dx.doi.org/10.1177/0022022102250225

Cheung, F. M. (2012). Mainstreaming culture in psychology. *American Psychologist, 67,* 721–730. http://dx.doi.org/10.1037/a0029876

Cheung, F. M., Leung, K., Zhang, J.-X., Sun, H.-F., Gan, Y.-Q., Song, W.-Z., & Xie, D. (2001). Indigenous Chinese personality constructs: Is the five-factor model complete? *Journal of Cross-Cultural Psychology, 32,* 407–433. http://dx.doi.org/10.1177/0022022101032004003

Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist, 66,* 593–603. http://dx.doi.org/10.1037/a0022389

Davidov, E. (2008). A cross-country and cross-time comparison of human values measurements with the second round of the European Social Survey. *Survey Research Methods, 2,* 33–46.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312. http://dx.doi.org/10.1037/1040-3590.6.4.304

Georgas, J. (1989). Changing family values in Greece: From collectivist to individualist. *Journal of Cross-Cultural Psychology, 20,* 80–91. http://dx.doi.org/10.1177/0022022189201005

Georgas, J., Berry, J. W., van de Vijver, F. J. R., Kaγitçibaşi, C., & Poortinga, Y. H. (Eds.). (2006). *Families across cultures: A 30-nation psychological study*. New York, NY: Cambridge.

Guo, X., Suarez-Morales, L., Schwartz, S. J., & Szapocznik, J. (2009). Some evidence for multidimensional biculturalism: Confirmatory factor analysis and measurement invariance analysis on the Bicultural Involvement Questionnaire-Short Version. *Psychological Assessment, 21,* 22–31. http://dx.doi.org/10.1037/a0014495

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Erlbaum.

Hambleton, R. K., & de Jong, J. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20,* 127–134. http://dx.doi.org/10.1191/0265532203lt247xx

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147–157. http://dx.doi.org/10.1027/1015-5759.11.3.147

Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46–74). New York, NY: Cambridge University Press.

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. P. (Eds.). (2003). *Cross-cultural survey methods*. Hoboken, NJ: Wiley.

He, J., & van de Vijver, F. J. R. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*(2). http://dx.doi.org/10.9707/2307-0919.1111

Hein, S., Reich, J., & Grigorenko, E. (2015). Cultural manifestation of intelligence in formal and informal learning environments during childhood. In L. A. Jansen (Ed.), *The Oxford handbook of human development and culture: An interdisciplinary perspective* (pp. 214–229). New York, NY: Oxford University Press.

Ho, D. Y. F. (1996). Filial piety and its psychological consequences. In M. H. Bond (Ed.), *Handbook of Chinese psychology* (pp. 155–165). New York, NY: Oxford University Press.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 4,* 179–188.

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

International Test Commission. (2005). International guidelines on test adaptation. Retrieved from http://www.intest.org/files/guidelines_test_adapt

Markus, H. R., & Kitayama, S. (1991). Cultural variation in the self-concept. In J. Strauss & G. R. Goethals (Eds.), *The self: Interdisciplinary approaches* (pp. 18–48). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4684-8264-5_2

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543. http://dx.doi.org/10.1007/BF02294825

Miller, K., Mont, D., Maitland, A., Altman, B., & Madans, J. (2011). Results of a cross-national structured cognitive interviewing protocol to test measures of disability. *Quality & Quantity: International Journal of Methodology, 45,* 801–815. http://dx.doi.org/10.1007/s11135-010-9370-4

Serpell, R. (2011). Social responsibility as a dimension of intelligence, and as an educational goal: Insights from programmatic research in an African Society. *Child Development Perspectives, 5,* 126–133. http://dx.doi.org/10.1111/j.1750-8606.2011.00167.x

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating guidelines for test adaptation: A methodological analysis of translation quality. *Journal of Cross-Cultural Psychology, 37,* 557–567. http://dx.doi.org/10.1177/0022022106290478

Valchev, V. H., Nel, J. A., van de Vijver, F. J. R., Meiring, D., de Bruin, G. P., & Rothmann, S. (2013). Similarities and differences in implicit personality concepts across ethnocultural groups in South Africa. *Journal of Cross-Cultural Psychology, 44,* 365–388. http://dx.doi.org/10.1177/0022022112443856

van de Vijver, F. J. R. (2009, July). *Translating and adapting psychological tests for large scale projects*. Paper presented at the 11th European Congress of Psychology, Oslo, Norway.

van de Vijver, F. J. R. (2011). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 3–34). New York, NY: Routledge/Taylor & Francis.

van de Vijver, F. J. R., Chasiotis, A., & Breugelmans, S. M. (2011). Fundamental questions of cross-cultural psychology. In F. J. R. van de Vijver, A. Chasiotis, & S. M. Breugelmans (Eds.), *Fundamental questions in cross-cultural psychology* (pp. 9–34). New York, NY: Cambridge.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1,* 89–99. http://dx.doi.org/10.1027/1016-9040.1.2.89

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17–45). New York, NY: Cambridge.

van de Vijver, F. J. R., & Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Mahwah, NJ: Erlbaum.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology / Revue Européenne de Psychologie Appliquée, 47,* 263–279.

van Leest, P. E. (1997). Bias and equivalence research in The Netherlands. *European Review of Applied Psychology / Revue Européenne de Psychologie Appliquée, 47,* 319–329.

Received January 13, 2015
Revision received July 13, 2015
Accepted September 12, 2015 ■