INSEEC Research

PARIS • BORDEAUX • LYON • ALPES-SAVOIE
MONACO • LONDON • CHICAGO

Working Paper
n°: 2013-17-01

Towards a Procedure for Developing Measurement Scales for Cross-Cultural Management Research

Marco Caramelli[a] & Fons van de Vijver[b]
[a] INSEEC Business School, 27 avenue Claude Vellefaux 75010 Paris, France
[b] Tilburg University, the Netherlands, and North-West University, South Africa

February 2011

An ulterior version of this article appeared in International Management Review (2013), Vol.17, n°2, pp. 150-163 - ISSN : 1206-1697 - DOI : 10.7202/1015406ar

It can be purchased at: http://www.managementinternational.ca/catalog/towards-acomprehensive-procedure-for-developing-measurement-scales-for-cross-culturalmanagement-research.html

Towards a Comprehensive Procedure for Developing Measurement Scales for Cross-Cultural Management Research

ABSTRACT

A procedure for developing and testing measurement scales for use in cross-cultural comparative management research is described. The procedure emphasizes the combination of adequate instrument design if a new instrument is used or adequate adaptation procedures if working with an existing instrument and sophisticated statistical analyses (multigroup confirmatory factor analysis) to test the adequacy of the scales in all groups. The procedure is illustrated in a study of "competitive orientation" among French, Mexican, US and Italian employees of multinational corporations.

Comparative organizational research consists of the systematic detection, identification, measurement and interpretation of similarities and differences of organizational behavior among employees of different cultural groups (Adler, 1983; Boddewyn, 1965). During the past decades, there has been a growing body of literature addressing the specific methodological problems of this type of research, such as the equivalence of constructs, samples and measurement instruments. Meaningful cross-group comparisons presuppose that the measurement instruments used to assess attitudes, values or behaviors, operate in an equivalent way across groups (i.e., that they measure the same thing in the same way). Otherwise, differences in mean levels or in the pattern of correlation of the measures are potentially artifactual and may be substantively misleading (Mullen, 1995; Raju, Byrne, & Laffitte, 2002; Reise, Widaman, & Pugh, 1993). Cross-national studies are certainly the most threatened by measurement invariance issues because concepts and measures designed for one country are applied to employees of a second country without modifications (Byrne & Campbell, 1999; Lim & Firkola, 2000; Schaffer & Riordan, 2003), and because translation is a main source of measurement non-invariance (B.M. Byrne & Watkins, 2003). However, authors have stressed the relevance of assessing measurement invariance in a broad variety of comparisons, such as time (longitudinal studies), modes of survey administration (e.g., online versus paper-and-pencil), employees of different organizational levels, gender, age, and race (Meade, 2010; Meade & Lautenschlager, 2004a; Vandenberg & Lance, 2000). Basically, this suggests that measurement equivalence should be ensured prior to *any* cross-group comparison (Schaffer & Riordan, 2003; Vandenberg & Lance, 2000). Issues regarding measurement equivalence are getting more and more popular in organizational research after the publication of several state-of-the art articles on the topic (e.g. Cavusgil & Das, 1997; Hui & Triandis, 1985; A. W. Meade & Lautenschlager, 2004a; Peng, Peterson, & Shyi, 1991;

Reise et al., 1993; Schaffer & Riordan, 2003; Singh, 1995; Steenkamp & Baumgartner, 1998; Vandenberg, 2002; Vandenberg & Lance, 2000). If such post hoc analyses give evidence of a severe lack of measurement equivalence, substantive comparisons cannot be performed, possibly even forcing the researcher to collect new data (Vandenberg, 2002). It is an important limitation of this statistical approach to equivalence that the quality check is only conducted post hoc and that little attempt is made to choose, adapt or develop instruments that maximize the likelihood of finding equivalence.

In this article, we suggest that it is important to incorporate equivalence issues in the scale development process. Our aim is to describe a step-by-step procedure for developing measures that are more likely to provide comparable scores in cross-group comparisons. All the classical steps of scale development procedures are addressed from the perspective of ensuring the equivalence of the concept to be measured; particular recommendations on how to deal with equivalence problems are discussed.

From an extensive review of published articles and books on cross-cultural methods in the last thirty five years, we could not find any other attempt to integrate the knowledge in cross-cultural methods with scale development procedures. More precisely, we extend previous research in three ways. First, we integrate measurement equivalence issues in each step of classical scale development procedures whereas previous articles described only some of the steps implied in scale development and validation. Second, we suggest that equivalence issues should be addressed on theoretical grounds, whereas existing research often proceeded in an “atheoretical” fashion. Third, we explain how to deal with equivalence problems in an accessible way, whereas the existing literature focuses more on the identification of these problems.

In the first section, we briefly present the conceptual background of cross-cultural measurement equivalence. In the second section, we describe a scale development procedure

that can be useful both in the context of a derived-etic approach (the researcher needs to develop a measurement scale in the context of a comparative study) and in the context of an etic approach (the researcher decides to use an existing scale for a cross-group comparison and wants to assess whether the scale is suitable for all the groups under study).

Conceptual Background

Validity, Reliability, and Measurement Equivalence

Measurement is a vital concern for all researchers in social sciences. The quality of a measure is generally assessed by its validity and reliability. The validity of a measure has been defined as "the extent to which an operationalization measures the concept it is supposed to measure" (Bagozzi, Yi, & Phillips, 1991: 421), whereas reliability is the proportion of variance of the obtained score attributable to the true score of a latent variable (De Vellis, 2003). Hence, the goal of our procedure is to develop measures which are equally valid and reliable for all the populations under study. This is usually called measurement equivalence (or invariance) between groups. Two measures applied to different groups are equivalent if the relationships between the observed score and the true score are identical for all groups. This essentially implies that the concept involves the same domain in all groups (true scores), and that the items composing the scale are equally appropriate in measuring the construct (observed scores).

When is Measurement Equivalence an Issue?

In which cases are group comparisons most likely to be threatened by a lack of measurement equivalence? Existing articles on measurement equivalence do not clearly address this issue. Vandenberg and Lance (2000) suggest that measurement equivalence tests should be performed when comparing different "populations" or "groups", whereas other authors refer to "cultural groups", "cultural samples" (e.g. Schaffer & Riordan, 2003; Vandenberg, 2002), "cross-national" groups (Steenkamp & Baumgartner, 1998) or "ethnic"

groups within countries (Van de Vijver, 1998). As suggested by Meade and colleagues (Meade, Michels, & Lautenschlager, 2007), tests of measurement equivalence have often proceeded in an atheoretical fashion in that between-group comparisons are made without a priori notions of whether equivalence would exist.

The practically important question is when equivalence issues have to be addressed in any of the distinctions described. We argue that equivalence should be addressed when differences in meaning of scores can be expected across the groups under study. It is not so much the expectation of significant differences in mean scores that is relevant, but the expectation that scores could have a different meaning should lead to concerns about equivalence. Put differently, cultural groups have traditionally been described in terms of their position (low vs. high) on a series of cultural dimensions (Hofstede, 2001; Lytle, Brett, Barsness, Tinsley, & Janssens, 1995; Morden, 1999). What we suggest here, is that cultural groups can be expected to vary on a certain construct (e.g., attitude towards participative management), when they are different in terms of a cultural dimension (for example power distance) likely to influence that construct (Lytle et al., 1995).

As an example, Wasti and colleagues (Wasti, Bergman, Glomb, & Drasgow, 2000) tested the generalizability of a model of the antecedents and consequences of sexual harassment developed in the US and applied to Turkish employees. The authors considered that tests of structural equivalence were necessary because Turkey is notably different in its *"*cultural, political, and economic orientation toward women, as compared with the US*"* which could have implications for the adequacy of the measures (Wasti et al., 2000: 767). In the case of a comparison between Swedish and Norwegian employees, measurement equivalence would be of less concern because of the strong similarity between these two countries concerning women issues.

Etics, Emics, and Derived-Etics

Three different research approaches have typically been used in cross-cultural organizational research to measure concepts and deal with equivalence. Most frequently (94% of the studies reviewed by Schaffer & Riordan in 2003), researchers start by using a concept and/or instrument developed within the frame of reference of one specific country (the US or another western country). Scales are subsequently translated and assumed to be a valid basis for comparison with other countries (Berry, 1989; Harkness, Van de Vijver, & Johnson, 2003). This *etic* (or imposed-etic) approach is the most widely used because it requires the fewest financial and time resources (Schaffer & Riordan, 2003) and because most researchers seek to produce generalizations across the cultural groups under study (Berry, 1989). The etic approach has been criticized on conceptual grounds in that a construct and its operationalization may not be valid for another group, which can lead to misleading comparisons.

The *emic* approach, on the other hand, attempts to define and operationalize a phenomenon occurring in a particular culture utilizing only insights from that culture. A researcher adopting an emic approach may obtain a very accurate within-culture description and insight but can easily run into equivalence problems when emically obtained data are compared across cultures (Davidson, Jaccard, Triandis, Morales, & Diaz-Guerrero, 1976). For example, a researcher can ask for culture-specific indicators of customer satisfaction in different countries. There is a fair chance that such emically developed instruments lack cross-cultural comparability because of cultural differences in what satisfies customers in different groups (e.g. Laroche, Ueltschy, Abe, Cleveland, & Yannopoulos, 2004).

Because of the drawbacks of both emic and etic approaches, researchers increasingly consider a *derived-etic* approach (or combined emic-etic approach; Cheung, Van de Vijver, & Leong, 2011) as a best practice that offers scope for both universal and culture-specific aspects of measures (Schaffer & Riordan, 2003). Such an approach amounts to first attaining

emic knowledge about all the cultures in the study and then retaining the communality as the basis for comparisons (Berry, 1989). The main strength of this approach is the attention for both ecological validity (by designing the measures on the basis of locally obtained information) and cross-cultural comparability. In this recommended approach, the quest for equivalence starts from the beginning of the research process and not right after data are collected (Usunier, 1998).

Scale Development Procedure

Step 1. Specify the Domain of the Construct – Construct Equivalence

The first important step in a process of measurement development is to determine clearly what one wishes to measure: "The researcher must be exacting in delineating what is included in the definition and what is excluded" (Churchill, 1979: 67). It is important to establish whether the construct exists in all groups and if the core and boundaries of the phenomenon are the same (Cavusgil & Das, 1997). Unfortunately, most researchers tend to address this issue only in a post hoc fashion at the stage of measurement invariance analysis (Hambleton, 2001; Peng, Peterson, & Shyi, 1991; Singh, 1995). We suggest that construct equivalence should be addressed in the first stages of the research process. Construct equivalence (also referred to as conceptual/functional equivalence) concerns the similarity of the definition of the concept, but also the similarity of the determinants, consequences, and correlates (Hui & Triandis, 1985). For example, conceptually equivalent behaviors or work attitudes may have developed in different cultures in response to similar problems (Sekaran, 1983). Various procedures can be used, such as an inspection of the literature, collaboration with colleagues from other countries, interviews and focus groups with individuals from the groups under study, to find out whether the concept exists in all the subgroups and which components are universal and culture-specific.

Step 2. Generate Items – Cultural Equivalence

The items of a scale should adequately cover the domain of the concept. In single-group research, construct coverage can be established by open-ended interviews with representative subjects from the target population so that items can be derived from the transcripts of the interviews, thereby ensuring the natural context and word choice (Churchill, 1979; Dawis, 1987). Reviews of existing literature on the concept under study and of existing measures are other often-used methods. In the latter case, items may have to be modified or rewritten to ensure cultural adequacy and to establish consistency in tone and perspective across all items (Bolino & Turnley, 1999).

In cross-cultural research the same process should be conducted for each group separately and potential indicators be listed. In a second step, common indicators will be selected in order to ensure that only items that are good indicators for all groups are retained. In order to maximize equivalence, it might be necessary to remove the cultural particulars. Imagine the case of a researcher who wants to develop a measure of the importance of prestige in several countries with large differences in economic development. In a poor country “Owning a watch” could be considered as an indicator of prestige which would not be the case in rich countries. Therefore, the use of more generic items such as “Success in my work is important for me” should be preferred because even if success can be conceived differently by different groups, it is more likely to be a good indicator of prestige than owning a watch.

Items of the initial pool may have to be reformulated to maximize their translatability. Brislin (1986) described guidelines which purpose is to ensure that the translators will clearly understand the meaning of the original language item, to have a high probability of finding a readily available target language equivalent, and to produce readily understandable target language items. Examples of such guidelines are to use simple, short sentences, to employ the active rather than the passive voice, and to repeat nouns instead of using pronouns. Even if

existing tests are employed, items may require modifications so as to maximize their adequacy in the new cultural groups or new items may need to be added to tap into additional aspects of a phenomenon not covered by the original test (Brislin, 1986).

Step 3. Translate and/or Adapt Items – Linguistic Equivalence

This step concerns mainly but not exclusively studies involving multiple languages. Even when a survey instrument is administered to different groups using the same language (e.g., English), it is important to ascertain that the vocabulary and the language style are familiar to all groups. For instance, De Vellis (2003) reports examples showing that the same word can have different meanings in different English-speaking countries and even within the same country.

Translation is one of the most frequently mentioned problems in the literature dealing with empirical comparative research (Harkness, 2003). Translation equivalence may be divided into different categories (Usunier, 1998). Evidence for lexical equivalence or similarity of denotation, is provided by dictionaries. Grammatical-syntactical equivalence deals with original and translated text similarities and how word order and other grammatical features are used to convey meaning. Finally, experiential equivalence is about what words and sentences mean for people in their everyday experience. For example, “manger des pâtes” (eating pasta) has a completely opposite affective meaning in France and in Italy. While in France this means having a cheap meal, in Italy, “mangiare la pasta” represents more the idea of a good meal.

Different procedures have been proposed to translate instruments. We discuss here the most common ones. The most widely used method is the back-translation technique (Harkness, 2003; Usunier, 1998). One bilingual translates from the source to the target language, and another blindly translates back to the source. The accuracy of the translation is evaluated by comparing the original and backtranslated versions. Nontrivial differences

between the versions are seen as evidence of translation problems. The procedure can be iteratively repeated for several rounds and a final target-language questionnaire is discussed and prepared by the researcher and the two translators (Brislin, 1986; Usunier, 1998; Van de Vijver & Leung, 1997). The back-translation technique has many advantages: it is less likely that the preliminary version is "contaminated" by one single person and no language is the center of attention (Brislin, 1986). However, some limitations of the back-translation technique have been observed. For instance, "good" back translators might automatically compensate for poorly translated texts and thus mask problems (Brislin, 1986; Harkness, 2003). Also, Van de Vijver and Leung (1997) point out that the procedure can produce a stilted language that does not approach the naturalness of the text in the original version. As a consequence, the use of back translation seems to be less and less recommended by psychometric experts (Byrne & Campbell, 1999). A second technique, called blind parallel translation, consists of having several translators translate independently from the source language into the target language. The different target versions are then discussed and a final version is compiled (Harkness, 2003; Usunier, 1998). In cross-cultural research, we often need to develop surveys in more than two languages. We propose here a modified version of the blind parallel translation technique that may be better suited for such multilingual studies. Suppose that a study involves English, Italian, and Spanish participants and that the mother tongue of the principal investigator is English. A first step involves two bilingual English-Spanish translators and two English-Italian translators proposing a translation of the first English version separately. After this step, we have two Spanish and two Italian versions of the questionnaire. The two pairs of translators then compare and discuss their translations until they agree on a common version.

Particular attention should be paid in the translation process to equivalence of response formats, because inadequate translation will lead to systematic cross-cultural differences.

Likert scales are the most widely used response scale in organizational research (Hinkin, 1995). Existing evidence shows that difficulties can occur in determining lexical equivalents in different languages of verbal descriptions for the scale and that it is difficult to ensure that the distances between scale points are equivalent in all the languages (Usunier, 1998). Several solutions have been proposed to increase the equivalence of response scales; these include the substitution of verbal intervals with numerical scales and the use of local wordings based on scales developed by local researchers (Smith, 2003; Usunier, 1998). The first option seems preferable because numbers are more likely to operate equivalently than words.

Step 4: Adjudication of Judgmental Aspects

The aim of this step is to assess the quality of the previous stages, and to improve the scale's face and content validity. When the constructs to measure are based on well-tested theory, the most widely used method consists of asking a group of experts to review the item pool (De Vellis, 2003; Hardesty & Bearden, 2004). These experts can provide information about the adequacy of the items; furthermore, if a scale is translated, bilingual experts cannot only compare the semantic similarity of the original and translated versions, but can also evaluate other text features such as comprehensibility.

Step 5: Collect Pilot Data

After an initial set of items for each group has been established, a pilot test is necessary (Churchill, 1979). It is generally recommended to use development samples that are sufficiently large and drawn from the target populations (De Vellis, 2003). Sample size is important for subsequent exploratory and confirmatory factor analyses. Being aware that more is always better, we can recommend a ratio of 10 to 20 people per measured variable, with 100 respondents per group being a bare minimum sample size (A. W. Meade & Lautenschlager, 2004b; Thompson, 2004).

The researcher should try to maximize the equivalence in data collection and avoid method bias which encompasses three aspects: sample bias relates to the comparability of samples; instrument bias derives from the responses to the format of the assessment instrument such as response sets and social desirability; finally, administration bias results from differential administration conditions such as interviewer effects (B.M. Byrne & Watkins, 2003).

Step 6: Assessment of Psychometric Properties in Each Sample

After data are collected, the validity and reliability of the scales should be evaluated for each group to ensure that appropriate items are retained to constitute the scales (Churchill, 1979; De Vellis, 2003). In single-group research, tests of validity and reliability typically start with an exploratory factor analysis to test whether the selected items show the hypothesized factor structure (Hair Jr., Black, Babin, Anderson, & Tatham, 2006). Scales are formed by taking all items together that load at least moderately on the same factor (e.g., having a standardized loading with an absolute value of at least .4) and do not load as high on other factors (Gerbing & Anderson, 1996; Hair Jr. et al., 2006). Items that do not show this convergent and discriminant validity are usually dropped (Campbell & Fiske, 1959). The validation phase typically ends with a confirmatory factor analysis to establish the final version of the scales (Hinkin, 1995).

It is generally recommended to collect new data so that the exploratory and confirmatory factor analyses are based on different data (Churchill, 1979; De Vellis, 2003). However, because of the difficulty of data collection in organizational settings, authors usually split their sample into two halves and perform exploratory analyses on the first half and confirmatory analyses on the second half. When validity is established, reliability is generally assessed by computing coefficient alpha. Again, authors usually delete some poorly performing items to increase coefficient alpha (Hinkin, 1995). Different norms about

minimum values for alpha have been proposed. A value of .70 is often taken as the minimal value that is required and values of at least .80 are seen as high (Cicchetti, 1994).

By factor analyzing the items for each group separately, one can check whether the same factors appear and if the items load on the same factors for all groups. Items failing to show cross-group convergent and discriminant validities can be eliminated from the cross-cultural comparison; yet, these may contain interesting information about cross-cultural differences in that they show very different patterns of loadings.

Note that in single-group research, EFA is generally only conducted when a measurement scale is created, and only CFA is conducted when scales from prior research are used. In cross-cultural research, it may be preferable to start with an EFA in both cases since scales from previous research are developed in the context of a single group and applied in the context of other groups.

Step 7: Assessment of the Equivalence of the Psychometric Properties – Measurement Equivalence

Cross-cultural research usually deals with either between-group comparisons of the latent or observed means of some concept (Type I or level-oriented studies; e.g., is work motivation higher in Japan than in China?) or with between-group comparisons of the relationships between constructs (Type II or structure-oriented studies; e.g., is the relationship between stock-options and motivation the same for top and middle managers?). The requirements in terms of measurement equivalence are different for these two types of comparisons. If differences in score levels are of interest, comparisons are only meaningful if the measurement scales have the same origin (zero point) and the same metric (scale units). If the issue of interest involves the relationships between two or more variables, the only requirement for meaningful comparisons is that the scale on which the scores are expressed have the same metric (Mavondo, Gabbott, & Tsarenko, 2003; Poortinga, 1989).

There is wide agreement in the management and organizational literature that multigroup confirmatory factor analysis represents the most powerful and versatile approach of testing for cross-cultural measurement invariance (Steenkamp & Baumgartner, 1998).

The procedure consists of testing the equivalence of the parameters of the measurement model as defined within the confirmatory factor analysis framework (Vandenberg & Lance, 2000). More specifically, the procedure consists of testing the goodness of fit of increasingly restrictive models. Models are nested meaning that placing equality constraints on one of the models produces the other (nested) model.

The measurement model should fit the data within any of the groups under scrutiny and, in case of a good fit, cross-group equivalence should be simultaneously assessed for all the groups. To date, authors do not completely agree on (1) which tests of measurement equivalence should be undertaken, (2) the sequence of the tests, (3) the substantive meaning of each level of invariance and (4) the extent to which partial equivalence can be accepted (see for example Vandenberg, 2002). In what follows, we present a synthesis of the main principles concerning measurement invariance as they are typically described in the literature, and we propose a way of dealing with the four issues mentioned above. We also go beyond past research in terms of interpretation of substantive results in the light of the results of the measurement invariance tests.

Researchers used to first perform an “omnibus test” of the equality of the covariance matrices across groups. If covariance matrices did not differ across groups, full measurement equivalence was considered to be established (Vandenberg & Lance, 2000). However, some authors have questioned the usefulness of this particular test on the grounds that it can indicate that measurement equivalence is supported when more specific tests of measurement equivalence find otherwise (A. W. Meade & Lautenschlager, 2004a; Raju et al., 2002). Then, it seems more reasonable to directly inspect each level of equivalence.

The first generally advocated test is the test of *configural invariance*, or whether respondents of different groups associate the same subsets of items with the same construct(s), meaning that the underlying cognitive domains are the same (Riordan & Vandenberg, 1994). The absence of nonnegligible difference in the pattern of fixed and free factor loadings between groups is usually taken as supportive and sufficient evidence of configural equivalence (Vandenberg, 2002). Configural invariance is a precondition for higher levels of measurement equivalence. Therefore, it is generally viewed as a baseline model against which further tests, based on more restrictive models, are evaluated.

If steps 1 to 6 of the presented procedure have been followed, *configural equivalence* would be expected. The implications of not finding configural invariance vary depending on how many items lack invariance. If only a limited number of items do not load on the specified factor in one or some groups, and there are still enough invariant items left, removing the items from the cross-cultural comparison may be desirable. However, it is important to determine whether the remaining common items still adequately cover the construct as defined at the beginning of the study or the original construct has to be narrowed.

The presence of anomalous stimuli indicates that some type of cultural specificity has been observed; for example, certain stimuli may measure secondary constructs or the content of the stimuli could be inappropriate in some cultures (e.g., the attitude towards living with one's parents may be an indicator of individualism in France but not in Italy or Spain where this refers more to a national norm). If, on the other hand, the lack of configural invariance is a consequence of many anomalous stimuli and the factor structures turn out to be essentially different in different groups, the implications are more severe. Such an observation means that the concept or various indicators are culture specific and that quantitative comparisons between groups, involving these indicators, are not meaningful.

Most authors advocate testing *metric equivalence* after configural invariance has been established. It involves the equality of scale units between groups and is required to compare relationships between variables in different groups. Metric equivalence concerns the relationship between the latent variable and its indicators and is tested by constraining the items' factor loadings to be invariant across groups (Steenkamp & Baumgartner, 1998).

The results of the analysis can point to full metric equivalence, complete absence of metric equivalence, and partial metric equivalence (at least one but not all item failed to be invariant). Since exact measurement invariance is unrealistic in many cases, an important question is what to do if metric invariance for all the items is not established (Labouvie & Ruetsch, 1995; Raju et al., 2002).

Depending on the outcome, three aspects have to be considered: the number of items that have different loadings, the size of the loading differences and, more specifically, whether the differences are large enough to be consequential in terms of cross-cultural substantive comparisons, and the size of the observed relationships between the latent variables under study. Concerning the number of items, it is technically possible to compare relations between constructs as soon as at least one indicator has invariant loadings. Practically speaking, however, if only very few items have invariant loadings, such item-specific comparisons do not convey much information about the underlying construct. Authors do not agree on how many invariant items are needed to accept partial metric equivalence (Cheung & Rensvold, 2002; Schaffer & Riordan, 2003; Vandenberg, 2002). In business and organizational research, short, unidimensional scales (< 10 items) are common (Hinkin, 1995). For such scales we propose to avoid comparisons of correlations between constructs if less than half of the items are metrically equivalent.

The second relevant issue in analyzing partial invariance is the size of the differences of the noninvariant loadings. There is no widely accepted rule as to when loadings are

sufficiently different to be psychologically consequential. For example, if an item loading for group A is .65 and for group B is .75, the difference can be statistically significant given a sufficiently large sample size but the difference is, psychologically speaking, very small.

According to Meade and Bauer (2007) when metric equivalence is not found, researchers can calculate effect sizes and confidence intervals for the factor loading differences. If these intervals are small and close to (but exclude) zero, then the difference in loadings is so small that substantive comparisons are still justified.

The degree of metric invariance should also be interpreted in the light of the strength of the observed correlations between the latent constructs of interest. For example, if differences in correlations between two variables (e.g. motivation and stock options) between two groups (low vs high level employees) are small, a lack of metric equivalence would have more serious implications than in the case of a strong difference of correlations.

A test of *scalar invariance* is the most currently used after metric invariance is established (Vandenberg & Lance, 2000). *Scalar invariance* concerns the consistency of the differences between latent means and observed means across groups. Even if an item measures the latent variable with the same metrics for different groups, scores on that item can still be systematically upward or downward biased (Steenkamp & Baumgartner, 1998). Scalar invariance is tested by imposing an equality constraint on the item intercepts. When testing for equivalence of item intercepts, the parameters that have been found to be inequivalent in previous analyses should be freely estimated. Only the item loadings found to be invariant and the item intercepts must be constrained to be equal between the groups.

The same three outcomes can be found (complete support, partial support, and complete lack of scalar equivalence). Again, quality and quantity have to be taken into account. If the measurement intercepts of different items consistently point to deviances in one or the same small sets of groups, it is important to check for consistencies of the bias across groups. If the

bias is not consistently pointing to a single group, it is more likely that the bias is due to item specific issues like inadequate translations. If there is more consistency, then systematic sources of bias, such as social desirability or acquiescence, may play a role.

If the test of scalar equivalence indicates that differences in item intercepts are not consistent across items, anomalies can be examined by removing the items from the cross-cultural comparison. Another approach is to evaluate the influence of the biased items on the cross-cultural differences; a simple way is to compare the difference in size of the means on the original instruments with the difference on the instruments from which all biased items are removed. Significances or effect sizes of the differences can be compared. Although the differences can be very large in theory, it is rather common to find that the removal of biased items does not have major implications for their size and that the implications for the interpretation of the cross-cultural differences are minimal (e.g., Meiring, Van de Vijver, & Rothmann, 2006). Finally, the size of the intercept differences and the cross-groups mean differences should be taken into account when interpreting lack of scalar equivalence. Again, if the mean difference of two variables between two groups is small, a lack of scalar equivalence would have more serious implications than in the case of a large difference of means.

Other tests can concern factor variance equivalence and error variance equivalence (Vandenberg & Lance, 2000). However, a consensus among researchers has developed that metric and scalar equivalence are sufficient for establishing that measurement equivalence conditions exist (A. W. Meade & Bauer, 2007).

A further issue concerns the selection of an item as the referent indicator for identification purposes (Hair Jr. et al., 2006). Typical practice is simply to select an item and fix the loading to the value of 1. However, this practice can lead to biased results if the researcher inadvertently selects as the referent indicator an item that is not metrically invariant

(Vandenberg, 2002). In order to avoid such an issue, researchers can inspect the item loadings from the EFA and select, as the referent indicator, the one whose loadings are the most similar across groups.

Note that tests of measurement equivalence based on item response theory (IRT) have also been proposed (Raju et al., 2002; Reise et al., 1993). However, IRT has mainly been used in the analysis of educational data and rarely in management research. This is partly due to the fact that most software for running IRT analyses such as Multilog, Winsteps and Difas, are not widely available in research institutions.

Example

The scale used in this example is a measure of competitive orientation, defined as the extent to which individuals perceive competition as something positive and desirable, which was used in a previous study performed by the first author. The sample comprises 368 French, 138 Mexican, 246 US and 276 Italian employees of 3 French multinational corporations. The survey was developed in English and then translated into French, Spanish, and Italian using the blind parallel translation procedure described above. Table 1 presents the US English version of this 5-item scale.

Insert Table 1 about here

Exploratory factor analysis and coefficient alpha. Principal axis factoring was used to test whether the selected items actually had the hypothesized factor structure for each group separately (Hair Jr. et al., 2006). For all groups, the scale was unidimensional and all items loaded in excess of .40 on the factor confirming convergent validity (Hair Jr. et al., 2006; Peter, 1981) (see Table 2). Reliabilities, as assessed by coefficient alpha, were of .80 for

France, .79 for Italy, .76 for the US, and .64 for Mexico, suggesting good reliabilities except for the group of Mexicans. Finally, a visual inspection of factor loadings suggested selecting item 5 as the referent indicator for CFA.

Insert Table 2 about here

Multigroup confirmatory factor analysis. Because chi-square-based tests of measurement invariance are highly sensitive to sample size (A. W. Meade & Bauer, 2007) we have extracted three random subsamples of 150 individuals for the French, Italian, and US groups, in order to compare groups of similar sample sizes. We used AMOS 7.0 (Arbuckle, 2005) to assess the validity of the hypothesized factor structure as well as measurement equivalence between French, Italian, US and Mexican employees. Three levels of invariance were assessed: *configural*, *metric*, and *scalar* invariance. As a prerequisite to testing for factorial equivalence, it is customary to assess for each group the measurement model which best represents the observed data. Afterwards, when testing for measurement invariance, equality constraints are imposed on particular parameters, and thus, data for all groups are analyzed simultaneously to obtain efficient estimates (B.M. Byrne, 2004). Table 3 shows the fit indices when the model is tested for each group separately. The model fitted poorly in the French, Italian, and US groups with all fit indices below the commonly recommended levels (Hair Jr. et al., 2006). An inspection of the modification indices suggested a strong correlation between the error variances of items 2 and 3 for the three groups. The fit indices improved strongly after including error covariances, leading to well fitting models for the four groups (models 1 to 7). It is important to note that model respecifications must be limited and

avoided when possible, because they may be driven by characteristics of the particular sample on which the model is tested (MacCallum, Roznowski, & Necowitz, 1992).

All standardized loadings were highly significant and ranged from .44 to .83. We then fitted the model for the four groups simultaneously to test for configural invariance and again we needed to estimate the error covariance between items 2 and 3 to yield good fit with a CFI value of .977 and a RMSEA value of .044 (model 9). Thus, it could be concluded that the Competition scale showed configural invariance across the four groups.

Insert Table 3 about here

The χ^2 value of 35.76 with 16 degrees of freedom provides the baseline value against which the subsequent tests of invariance are compared. Having established configural equivalence, the next logical test concerns metric equivalence (i.e., a test of invariant factor loadings). Model 10 in Table 3 shows the fit indices of the model where the factor loadings are constrained to be equal across groups. However, of primary importance is the comparison of its χ^2 value of 59.31 (df = 28) with that for the baseline model (χ^2 = 35.76, df = 16). In fact, when models are nested, this difference in χ^2 values (in large samples) follows a χ^2 distribution, with degrees of freedom equal to the difference in degrees of freedom (Van de Vijver & Leung, 1997). After the publication of the Cheung and Rensvold (2002) article, it has become common to use both the chi-square and the CFI differences (a drop in CFI that is smaller than or equal to .01 indicates that the null hypothesis of invariance should not be rejected) for evaluating model differences. This comparison yielded a χ^2 difference value of 23.55 with 12 *df*, which is significant ($p < .05$) and a drop of CFI of .013. Metric invariance was therefore rejected. Several strategies could then be implemented to determine which item loadings were non-invariant between which groups. We chose here to check whether metric

invariance (and subsequently scalar equivalence) held across groups taken two by two. The details of each test are reported in Table 3 (models 11 to 31). Overall, the results of our analyses of metric equivalence showed that the Competition scale was fully metrically invariant between the four groups. Item 4 for the Mexican group was somewhat different from the loadings of the other three groups, but the difference was not significant in terms of chi-square difference, but it was in terms of drop in CFI.

Insert Table 4 about here

As previously suggested, the interpretation of tests of measurement equivalence should also take into account the actual size of the parameter differences. As shown in Table 4, the factor loading's value for Items 4 was of .56 for the Mexican group compared to values of .77 for the United States group and .83 for the other two groups. In order to assess the size of the factor loading differences, we computed standard error and confidence intervals (CI) for the difference of factor loadings for Item 4, between France and Mexico (representing the largest difference with Italy). The CI for Item 4 for the group France was .68 - .99 and .43 – .71 for Mexico. Also, the standard error for the difference was of .209 and the 95% confidence intervals went from -.149 to +.67 suggesting that the difference between the loadings is small. Then, one may consider that partial metric equivalence is acceptable given that only one item out of five was found to be non-invariant (for only one group out of four). Concerning scalar equivalences, the results showed that all the item intercepts for the Mexican group were significantly higher than the other groups' intercepts. This can be interpreted as an acquiescence bias. Italian respondents showed systematically lower intercepts than the other groups (except for item 3), suggesting a nay-saying response style. Finally, we found partial

scalar equivalence when we compared the US and Italian groups, with scalar invariance not holding for Item 6.

Insert Table 5 about here

In order to interpret such results, it is important to take both the size of the intercept differences and the size of the scale means differences into account. The observed means for Competition are 3.17 for Italy, 3.27 for the US, 3.32 for France and 4.05 for Mexico on a five-point scale. Table 5 shows the mean differences and statistical significance for the four groups using the ANOVA post hoc multiple comparisons routine with SPSS 18. A first striking result concerns the Mexican group whose mean score is much higher than the others (from .73 to 88). In order to interpret the size of the intergroup intercept differences shown in Table 4, we computed confidence intervals for the differences between Mexico and Italy (the group with the lowest intercepts). The result for item 4 (the item with the largest difference of intercepts) was a lower bound of .76 and an upper bound of 1.17 for a 95% confidence interval which is quite low. Then, since the mean difference is large and the difference in intercepts is low, we can safely consider that (in this sample) the Mexicans are higher in competition than the other groups. The second result of interest is the mean difference between the French and the Italians since we previously found a complete lack of scalar invariance between those two groups. The mean difference of .14 is not statistically significant (p =0.65) but can be considered as essentially noninterpretable given the lack of scalar invariance. There is a risk that this observed mean difference could be attributable to some extent to response style differences and it would be safer to conclude that we cannot unambiguously determine whether the French and the Italians differ in their level of competition. Concerning the remaining groups, we found partial scalar equivalence for the Italy/US comparison. The mean

difference between Italy and the US is of .10 and is not statistically significant. Because only one item out of five fails to show scalar equivalence, we can safely conclude that these two groups do not differ in competition.

Conclusion

Comparing the attitudes and behaviors of different groups of employees or consumers is one of the most common aims of organizational research. However, meaningful cross-group comparisons presuppose that the measurement instruments used to assess attitudes, values or behaviors in organizations, operate in an equivalent way across groups. Otherwise, differences in mean levels or in the pattern of correlation of the variables are potentially artifactual and may be substantively misleading. In the organizational literature, the issue of measurement equivalence has become increasingly popular in particular when the groups being compared are composed of individuals from different countries. We have argued here that the issue of measurement equivalence should be addressed for any group comparison, when there are some reasons to expect between-group differences in the existence and definition of the constructs themselves, and in the capacity of a set of items to cover the domain of the constructs in an equivalent way. Moreover, based on an extensive review of the literature on the topic, we argued that researchers tend to address measurement equivalence in a post hoc fashion (that is, after data were collected). In this article, we suggested that researchers should start incorporating equivalence issues from the scale development process to increase the likelihood of getting equivalent measures. We then integrated measurement equivalence issues in each step of classical procedures of scale development to propose a step-by-step procedure of scale development for comparative research which would be useful for both researchers who need to develop a measurement scale and for researchers who want to use an existing scale in the context of a comparative study and want to assess whether the scale is suitable for all the groups under study. In this procedure, we described how a

combination of instrument design or adaptations and sophisticated statistical analyses can go a long way to enhance the validity of substantive results in comparative studies. We also stressed the importance of combining qualitative and quantitative methods as well as etic and emic approaches.

We finally presented an example of measurement equivalence analysis based on multigroup confirmatory factor analysis, and went beyond past research by suggesting how to deal with measurement non-invariance. As an example, we suggested that the degree of inequivalence between parameters as well as the strength of the substantive effects being studied (correlation between variables and mean differences) had to be analyzed. We believe that the overall framework presented in this article will help researchers dealing with complex issues in a straightforward and effective way.

References

Adler, N. J. (1983). "A typology of management studies involving culture." *Journal of International Business Studies,* Vol.14, No2, pp. 29-47.

Arbuckle, J. L. (2005). *Amos 7.0 user's guide*. Chicago, IL: SPSS.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). "Assessing construct validity in organizational research." *Administrative Science Quarterly,* Vol. 36, No.3, pp. 421-458.

Berry, J. W. (1989). "Imposed etics-emics-derived etics: The operationalization of a compelling idea." *International Journal of Psychology,* Vol. 24, No. 6, pp. 721-735.

Boddewyn, J. (1965). "The comparative approach to the study of business administration." *Academy of Management Journal,* Vol. 8, No.4, pp. 261-267.

Bolino, M. C., & Turnley, W. H. (1999). "Measuring impression management in organizations: A scale development based on the Jones and Pittman taxonomy." *Organizational Research Methods,* Vol. 2, No.2, pp. 187-206.

Brislin, R. W. (1986). “The wording and translation of research instruments.” In W. J. Lonner & J. W. Berry (Eds.), *Fields methods in cross-cultural research* (Vol. 8, pp. 291-324). Beverly Hills, CA: Sage.

Byrne, B. M. (2004). “Testing for multigroup invariance using AMOS graphics : A road less traveled.” *Structural Equations Modeling,* Vol.11, No. 4, pp. 272-300.

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,* Vol. 30, No. 5, pp. 555-574.

Byrne, B. M., & Watkins, D. (2003). “The issue of measurement invariance revisited.” *Journal of Cross-Cultural Psychology,* Vol. 34, No. 2, pp. 155-175.

Campbell, D. T., & Fiske, D. W. (1959). “Convergent and discriminant validation by the multitrait-multimethod matrix.” *Psychological Bulletin,* Vol. 56, pp. 81-105.

Cavusgil, S. T., & Das, A. (1997). “Methodological issues in empirical cross-cultural research: A survey of the management literature and a framework.” *Management International Review,* Vol. 37, No. 1, pp. 71-96.

Cheung, G. W., & Rensvold, R. B. (2002). “Evaluating goodness-of-fit indexes for testing measurement invariance.” *Structural Equation Modeling,* Vol. 9, No.2, pp. 233-255.

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*. Advance online publication.

Churchill, G. A. (1979). “A paradigm for developing better measures of marketing constructs.” *Journal of Marketing Research,* Vol. 16, No.1, pp. 64-73.

Cicchetti, D. V. (1994). “Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.” *Psychological Assessment,* Vol. 6, No.4, pp. 284-290.

Davidson, A. R., Jaccard, J. J., Triandis, H. C., Morales, M. L., & Diaz-Guerrero, R. (1976). « Cross-cultural model testing: Toward a solution of the etic-emic dilemma." *International Journal of Psychology,* Vol. 11, No.1, pp. 1-13.

Dawis, R. V. (1987). "Scale construction." *Journal of Counselling Psychology,* Vol. 34, No.4, pp. 481-489.

De Vellis, R. F. (2003). *Scale development, theory and applications* (Vol. 26). Thousand Oaks, CA: Sage.

Gerbing, D. W., & Anderson, J. C. (1996). "An updated paradigm for scale development incorporating unidimensionality and its assessment." *Journal of Marketing Research*, Vol. 25, No.5, pp. 186-192.

Hair Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson.

Hambleton, R. K. (2001). "The next generation of the ITC test translation and adaptation guidelines." *European Journal of Psychological Assessment,* Vol. 17, No.3, pp. 164-172.

Hardesty, D. M., & Bearden, W. O. (2004). "The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs." *Journal of Business Research,* Vol. 57, pp. 98-107.

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

Harkness, J., Van de Vijver, F. J. R., & Johnson, T. P. (2003). "Questionnaire design in comparative research." In J. A. Harkness, F. J. R. Van de Vijver & P. P. Mohler (Eds.), *Cross-cultural survey methods*. Hoboken, NJ: Wiley.

Hinkin, T. R. (1995). "A review of scale development practices in the study of organizations." *Journal of Management,* Vol. 21, No.5, pp. 967-988.

Hofstede, G. (2001). *Culture's consequences - comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Hui, C. H., & Triandis, H. C. (1985). "Measurement in cross-cultural psychology." *Journal of Cross-Cultural Psychology,* Vol. 16, No. 2, pp. 131-152.

Labouvie, E., & Ruetsch, C. (1995). "Testing for equivalence of measurement scales: Simple structure and metric invariance reconsidered." *Multivariate Behavioral Research,* Vol. 30, No.1, pp. 63-76

Laroche, M., Ueltschy, L. C., Abe, S., Cleveland, M., & Yannopoulos, P. (2004). "Service quality perceptions and customer satisfaction: Evaluating the role of culture." *Journal of International Marketing,* Vol. 12, No.3, pp. 58-85.

Lim, L., & Firkola, P. (2000). "Methodological issues in cross-cultural management research: Problems, solutions and proposals." *Asia Pacific Journal of Management,* Vol.17, pp. 133-154.

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

MacCallum, R., Roznowski, M., & Necowitz, L. (1992). "Model modifications in covariance structure analysis: The problem of capitalization on chance." *Psychological Bulletin,* Vol. 111, No.3, pp.490-504.

Mavondo, F., Gabbott, M., & Tsarenko, Y. (2003). "Measurement invariance of marketing instruments: An implication across countries." *Journal of Marketing Management,* Vol.19, No.5/6, pp. 523-540.

Meade, A. (2010). "A Taxonomy of Effect Size Measures for the Differential Functioning of Items and Scales." *Journal of Applied Psychology,* Vol. 95, No.4, pp. 728-743.

Meade, A., Michels, L., & Lautenschlager, G. (2007). "Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study." *Organizational Research Methods,* Vol. 10, No.2, pp. 322.

Meade, A. W., & Bauer, D. J. (2007). "Power and precision in confirmatory factor analyic tests of measurement invariance." *Structural Equation Modeling,* Vol. 14, No. 4, pp. 611-635.

Meade, A. W., & Lautenschlager, G. J. (2004a). "A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance." *Organizational Research Methods,* Vol. 7, No.4, pp. 361-388.

Meade, A. W., & Lautenschlager, G. J. (2004b). "A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance." *Structural Equation Modeling,* Vol. 11, No.1, pp. 60-72.

Morden, T. (1999). "Models of national culture - a management review." *Cross Cultural Management,* Vol. 6, No.1, pp. 19-44.

Mullen, M. R. (1995). "Diagnosing measurement equivalence in cross-national research." *Journal of International Business Studies,* Vol. 26, No.3, pp. 573-596.

Peng, T. K., Peterson, M. F., & Shyi, Y. P. (1991). "Quantitative methods in cross-national management research: Trends and equivalence issues." *Journal of Organizational Behavior,* Vol.12, No.2, pp. 87-107.

Peter, J. P. (1981). "Construct validity: A review of basic issues and marketing practices." *Journal of Marketing Research, 18,* 133-145. Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology,* Vol. 24, pp. 737-756.

Raju, N. S., Byrne, B. M., & Laffitte, L. J. (2002). “Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory.” *Journal of Applied Psychology,* 87, 517-529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). “Confirmatiry factor analysis and item response theory: Two approaches for exploring measurement invariance.” *Psychological Bulletin,* Vol.114, No.3, pp. 552-566.

Riordan, C. M., & Vandenberg, R. J. (1994). “A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in a equivalent manner?” *Journal of Management,* Vol. 20, No.3, pp. 643-673.

Schaffer, B. S., & Riordan, C. M. (2003). “A review of cross-cultural methodologies for organizational research: A best-practices approach.” *Organizational Research Methods,* Vol. 6, No.2, pp. 169-216.

Sekaran, U. (1983). “Methodological and theoretical issues and advancements in cross-cultural research.” *Journal of International Business Studies,* Vol. 14, No. 2, pp.61-74.

Singh, J. (1995). “Measurement issues in cross-national research.” *Journal of International Business Studies*, Vol. 26, No.3, pp. 597-619.

Smith, T. W. (2003). “Developing comparable questions in cross-national surveys”. In J. A. Harkness, F. Van de Vijver & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 69-92). Hoboken, NJ: Wiley.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). “Assessing measurement invariance in cross-national consumer research”. *The Journal of Consumer Research,* Vol. 25, No.1, pp. 78-90.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Society.

Usunier, J.-C. (1998). *International & cross-cultural management research*. Thousand Oaks, CA: Sage.

Van de Vijver, F. J. R. (1998). "Towards a theory of bias and equivalence." In J. Harkness (Ed.), *ZUMA-Nachrichten Spezial No.3. Cross-Cultural Survey Equivalence*. Mannheim, Germany: ZUMA.

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

Vandenberg, R. J. (2002). "Toward a further understanding of and improvement in measurement invariance methods and procedures." *Organizational Research Methods,* Vol. 5, No.2, pp. 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). "A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research." *Organizational Research Methods,* Vol. 3, No.1, pp. 4-70.

Wasti, S. A., Bergman, M. E., Glomb, T. M., & Drasgow, F. (2000). "Test of the generalizability of a model of sexual harassment." *Journal of Applied Psychology,* Vol.85, No.5, pp. 766-778.

Table 1

Scale of Competition (US English Version)

"*Indicate your degree of agreement or disagreement with the following statements:*" 1 = strongly agree, 5 = strongly disagree
(Item 1) Competition between employees usually does more harm than good (reverse score)
(Item 2) I enjoy working in situations involving competition with others
(Item 3) I believe I have a highly competitive spirit
(Item 4) When we compete against others, we give the best of ourselves
(Item 5) Competition makes us improve our skills

Table 2

Exploratory Factor Analysis and Coefficients Alpha

Items	France	Italy	United States	Mexico
1	.62	.68	.57	.42
2	.80	.81	.71	.74
3	.75	.66	.72	.67
4	.81	.80	.78	.69
5	.79	.79	.80	.79
Alpha	.80	.79	.76	.64

Table 3

Summary of Fit Statistics for Tests of Invariance

Model Description	χ^2	*df*	CFI	TLI	SRMS	RMSEA	ΔDF	$\Delta\chi^2$	ΔCFI
Monogroup Analyses									
1-France	25.93	*5*	.925	.85	.049	.15			
2-Covariance error items 2&3	11.52	*4*	.973	.93	.032	.10			
3-Italy	45.50	*5*	.850	.71	.076	.22			
4- Covariance error items 2&3	13.44	*4*	.970	.91	.042	.12			
5-United States	25.45	*5*	.902	.80	.064	.16			
6- Covariance error items 2&3	9.69	*4*	.982	.96	.035	.076			
7-Mexico	3.34	*5*	1	1	.026	0.00			
Multigroup Analyses									
(4 groups)									
8- Configural Equivalence	100.21	*20*	.907	.82	.026	.079			
9- Covariance error 2&3	35.76	*16*	.977	.94	.002	.044			
10- Metric Equivalence	59.31	*28*	.964	.95	.048	.042	12	23.55*	-.013

Model Description	χ^2	*df*	CFI	TLI	SRMS	RMSEA	ΔDF	$\Delta\chi^2$	ΔCFI
Bigroup Analyses									
France/Italy									
11-Configural Equivalence	24.97	8	.969	.92	.032	.079			
12-Metric Equivalence	31.24	12	.965	.94	.043	.069	4	6.27	-.004
13-Scalar Equivalence	56.71	17	.929	.92	.048	.083	5	25.47***	-.036
France/United States									
14-Configural Invariance	19.22	8	.977	.94	.031	.065			
15-Metric Invariance	28.09	12	.972	.94	.040	.058	4	8.87	-.005
16-Scalar Invariance	34.61	17	.964	.95	.040	.056	5	6.52	-.008
France/Mexico									
17- Configural Invariance	14.62	8	.983	.957	.032	.051			
18- Metric Invariance	23.23	12	.971	.951	.038	.054	4	8.61	-.012
19- Partial Metric Inv. (item4)	16.92	11	.981	.972	.069	.041	3	2.3	-.002
20- Scalar Equivalence	276.69	18	.328	.253	.136	.213	6	253.46***	-.64
United States/Italy									

Table 3 (continued)

Model Description	χ^2	*df*	CFI	TLI	SRMS	RMSEA	ΔDF	$\Delta\chi^2$	ΔCFI
21-Configural Invariance	24.98	9	.967	.926	.0354	.074			
22-Metric Invariance	28.96	12	.965	.941	.0461	.066	4	3.98	-.002
23-Scalar Invariance	44.98	17	.942	.932	.0481	.071	5	16.02**	-.023
24-Partial Scalar Inv. (Item 6)	34.56	16	.961	.952	.0465	.060	4	5.6	-.004
United States/Mexico									
25-Configural Invariance	10.79	8	.991	.978	.035	.034			
26-Metric Invariance	17.74	12	.981	.969	.038	.040	4	6.95	-.01
27-Scalar Invariance	119.99	16	.660	.581	.103	.146	5	105.48***	-.329
Mexico/Italy									
28-Configural Invariance	16.64	8	.978	.944	.0419	.059			
29- Metric Invariance	24.09	12	.966	.947	.0460	.057	4	7.45	-.012
30- Partial Metric Inv. (item4)	18.09	11	.979	.962	.0461	.048	3	1.45	-.001
31- Scalar Invariance	289.60	17	.286	.206	.1734	.222	6	265.51***	-.68

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

Table 4

Factor Loadings (λ) and Intercepts (τ) for France, Italy, the United States and Mexico

	France		Italy		United States		Mexico	
Items	λ	τ	λ	τ	λ	τ	λ	τ
Item 1	.50	2.82	.55	2.54	.44	2.78	.28	3.37
Item 2	.68	3.11	.61	2.97	.48	3.00	.64	3.92
Item 3	.56	3.36	.41	3.56	.53	3.47	.57	4.34
Item 4	.83	3.49	.83	3.32	.77	3.36	.56	4.30
Item 5	.74	3.79	.80	3.48	.82	3.75	.73	4.34

Table 5

Mean Differences for Competition

Group 1	Group 2	Difference	Significance
France	Italy	.14	.006
	United States	.04	.578
	Mexico	-.73	.000
Italy	United States	-.10	.205
	Mexico	-.88	.000
United States	Mexico	-.77	.000