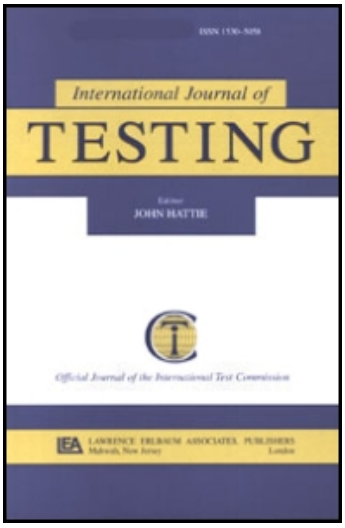
Testing for measurement and structural equivalence in large-scale cross-cultural studies

Byrne, B.M.; van de Vijver, F.J.R.

Published in:
International Journal of Testing

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record



International Journal of Testing
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653658

Testing for Measurement and Structural Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of Nonequivalence

Barbara M. Byrne[a]; Fons J. R. van de Vijver[bc]
[a] University of Ottawa, [b] Tilburg University, The Netherlands [c] North-West University, South Africa

Online publication date: 19 August 2010

To cite this Article Byrne, Barbara M. and van de Vijver, Fons 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: 2, 107 — 132
To link to this Article: DOI: 10.1080/15305051003637306
URL: http://dx.doi.org/10.1080/15305051003637306

International Journal of Testing, 10: 107–132, 2010

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305051003637306

Routledge
Taylor & Francis Group

Testing for Measurement and Structural Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of Nonequivalence

Barbara M. Byrne
University of Ottawa

Fons J. R. van de Vijver
Tilburg University, The Netherlands and North-West University, South Africa

A critical assumption in cross-cultural comparative research is that the instrument measures the same construct(s) in exactly the same way across all groups (i.e., the instrument is measurement and structurally equivalent). Structural equation modeling (SEM) procedures are commonly used in testing these assumptions of multigroup equivalence. However, when comparisons comprise large-scale cross-cultural studies, the standard SEM strategy can be extremely problematic both statistically and substantively. Based on responses to a 14-item version of the Family Values Scale (Georgas, 1999) by 5,482 university students from 27 nations around the globe, we describe and illustrate these difficulties. We propose and report on a dual modal two-pronged strategy that focuses on countries as well as scale items in determining the possible sources of bias. Suggestions for minimizing problems in tests for multigroup equivalence in large-scale cross-cultural studies are proffered.

Keywords: measurement equivalence, structural equation modeling, cross-cultural research, large-scale studies

In cross-cultural research, empirical investigations typically focus on tests for mean group differences. A critical assumption in such research, however, is that both the measuring instrument and the construct being measured are operating

Correspondence should be sent to Barbara M. Byrne. E-mail: bmbch@uottawa.ca

in the same way across the populations of interest. That is, there is presumed equality of (a) factorial structure (i.e., same number of factors and pattern of item loadings onto these factors), (b) perceived item content, (c) factor loadings (i.e., similar size of item estimates), and (d) item intercepts (i.e., item means). Given their psychometric focus, these characteristics are commonly regarded as representing *measurement equivalence* (also termed *measurement invariance*). Likewise, there is presumed equality of the measured construct with respect to (e) its dimensionality (i.e., unidimensional or multidimensional structure) and (f) in the case of multidimensional structure, relations among the construct dimensions. Given a focus on theoretical structure (see Bentler, 1978), the latter characteristics are considered to represent *structural equivalence* (also termed *structural invariance*). These assumptions, as is the case for all statistical assumptions, need to be tested. Indeed, Vandenberg and Lance (2000) have cautioned that failure to establish measurement and structural equivalence is as damaging to substantive interpretations as the inability to demonstrate reliability and validity. Fortunately, these equality assumptions are readily testable using structural equation modeling (SEM) procedures within the framework of a confirmatory factor analytic (CFA) model.

From reviews of the SEM literature (see Austin & Calderón, 1996; Hershberger, 2003; Tremblay & Gardner, 1996), it is evident that the past two decades have witnessed rapidly increasing use of SEM procedures in general and as they apply to tests for multigroup equivalence in particular (see Vandenberg & Lance, 2000). Possibly as a consequence of several pedagogical publications that have addressed issues of measurement and structural equivalence, as well as demonstrated the SEM approach to testing for these equivalencies, researchers today are more cognizant of this necessary prerequisite to testing for mean group differences. Thus, in addition to a rich bank of substantive research that has tested for equivalence across a variety of comparative groups (e.g., gender, age, organizations, academic track, teacher panels, culture, to name a few), there are many studies that have addressed technical issues associated with this procedure (such as partial measurement equivalence, effects of non-normality, and practical versus significant differences in comparative model fit).

One possible limitation of the standard SEM approach to testing for equivalence would appear to lie with its application to large-scale cross-cultural data, where confusing arrays of possibly nonequivalent parameters are often reported (e.g., Davidov, 2008; Welkenhuysen-Gybels, van de Vijver, & Cambré, 2007). The nature of the nonequivalence is often unclear. Is it due to conceptual misspecification that points to a serious lack of equivalence; is it due to accumulated, small, and inconsequential differences in parameters; or is it due to a combination of both?

The present article addresses this limitation. Specifically, the purposes are threefold: (a) to identify aspects of cross-cultural data that can seriously impact

attainment of multigroup equivalence; (b) to exemplify the extent to which use of the standard SEM approach to testing for multigroup equivalence can be problematic when several cultural groups are of interest; and (c) presented with findings of an ill-fitting baseline multigroup CFA model, to propose and illustrate a dual modal two-pronged approach to testing for multigroup equivalence that focuses on both countries and scale items as sources of possible bias.

Our review of the literature is presented in three sections. We begin by presenting an overview of the standard SEM approach to testing for multigroup equivalence in general and then, of its particular application to cross-cultural research. In Section 2, we outline complexities inherent in use of the standard approach to testing for equivalence when the data represent multicultural, rather than monocultural groups, and propose an alternate approach to these analyses. Finally, Section 3 identifies a critical limitation in using the standard approach to testing for equivalence across multicultural groups and outlines a modified approach that can address this limitation in large-scale cross-cultural studies.

TESTING FOR MULTIGROUP EQUIVALENCE

The General Notion

Development of a procedure capable of testing for multigroup equivalence derives from the seminal work of Jöreskog (1971). Although Jöreskog initially recommended that all tests of equivalence begin with a global test of equivalent covariance structures (i.e., variance-covariance matrices) across groups, this test has since been disputed because it often leads to contradictory findings. Exclusion of this omnibus test has therefore become common.[1] The classical approach to testing for factorial equivalence encompasses a series of hierarchical steps that begins with the determination of a well-fitting baseline multigroup model for which sets of parameters are put to the test of equality in a logically ordered and increasingly restrictive manner. (For a more detailed review and illustrated application of these steps, readers are referred to Byrne, 2006, 2008, 2009).

The first and least restrictive model to be tested is the baseline multigroup model noted above, which in SEM parlance is commonly termed the *configural* model (Horn & McArdle, 1992). With this initial model, only the extent to which the same number of factors and pattern (or configuration) of fixed and freely estimated parameters holds across groups is of interest and thus no equality constraints are imposed. In other words, for each group, the same model of hypothesized factorial structure is tested. The importance of the configural model is that it serves as the baseline against which all subsequent tests for equivalence are compared and thus, acceptable goodness-of-fit between this initial model and the multigroup data is imperative. In contrast, all remaining tests for equivalence involve the specification of cross-group equality constraints for particular parameters. The first

three constrained models test for measurement equivalence, while the remaining two test for structural equivalence. Measurement equivalence must be established prior to testing for structural equivalence.

Testing for measurement equivalence. This set of three tests always focuses on equality of the factor loadings across groups and can additionally include the item intercepts (i.e., the indicator or observed variable means) and their associated error uniquenesses,[2] respectively. Tests for the equivalence of factor loadings have been termed "metric equivalence" (Horn & McArdle, 1992)[3] as well as "measurement unit equivalence" (van de Vijver & Leung, 1997); tests for the equivalence of intercepts have been termed "scalar equivalence" (van de Vijver & Leung, 1997). Because each of these tests represents an increased level of restrictiveness, Meredith (1993) categorized them as weak, strong, and strict tests of equivalence, respectively.

In testing for the equivalence of factor loadings, these parameters are freely estimated for the first group only; this group is arbitrarily chosen and serves as the reference group. For all remaining groups, factor loading estimates are constrained equal to those of the reference group. Provided with evidence of equivalence, these factor loading parameters remain constrained as subsequent tests for the equivalence of additional parameters are conducted. On the other hand, given findings of nonequivalence related to particular factor loadings, one may proceed with subsequent tests for equivalence if the data meet the conditions of partial measurement equivalence noted by Byrne, Shavelson, and Muthén (1989). Although the specification of partial measurement equivalence has sparked a modest debate in the technical literature (see Millsap & Kwok, 2004; Widaman & Reise, 1997), its application remains a popular strategy in testing for multigroup equivalence, especially so in the area of cross-cultural research.

Tests for a well-fitting configural model and for invariant factor loadings are based on the analysis of covariance structures, which assumes that all observed variables (e.g., item scores, subscale scores) are measured as deviations from their means (i.e., their means are equal to zero). However, in moving on to the next more restrictive test of measurement equivalence, the equality of item intercepts, analyses are based on mean and covariance structures (i.e., item means are no longer zero); that is, analyses are based on the moment matrix, which includes both the sample means and covariances. Of import in testing for the equivalence of cross-group intercepts is that it subsequently allows for the multigroup comparison of latent construct means (i.e., means on the factors), should this be of interest. Although some researchers contend that this "strong" test of equivalence (i.e., test for invariant item intercepts) should always be conducted (e.g., Little, 1997; Little, Card, Slegers, & Ledford, 2007; Meredith, 1993; Selig, Card, & Little, 2008), others argue that analysis of only covariance structures may be the most

appropriate approach to take in addressing the issues and interests of a study (see, e.g., Cooke, Kosson, & Michie, 2001; Marsh, 1994, 2007; Marsh, Hau, Artelt, Baumert, & Peschar, 2006). Construct validity studies pertinent to a particular assessment scale (e.g., Byrne, Baron, & Balev, 1998) or theoretical construct (e.g., Byrne & Shavelson, 1986) exemplify such research.

The final and most stringent test of measurement equivalence (i.e., strict equivalence) focuses on the equality of error uniqueness variances across groups. It is now widely accepted that this test for equivalence is not only of least interest and importance (Bentler, 2005; Widaman & Reise, 1997) but also somewhat unreasonable (Little et al., 2007) and indeed not recommended (see Selig et al., 2008). One important exception to this widespread consensus, however, is in testing for multigroup equivalence of item reliability (see, e.g., Byrne, 1988).

Testing for structural equivalence. In contrast to tests for measurement equivalence, which focus on aspects of the observed variables, tests for structural equivalence center on the unobserved (or latent) variables. In the case of testing for the equivalence of a measuring instrument across groups, interest can focus on both the factor variances and their covariances, although the latter are typically of most interest. A review of the SEM literature reveals much inconsistency regarding whether researchers test for structural equivalence. In particular, these tests are of critical import to construct validity researchers whose interests lie either in testing the extent to which the dimensionality of a construct, as defined by theory, holds across groups (see, e.g., Byrne & Shavelson, 1987), or in the extent to which an assessment scale, developed within the framework of a particular theory, yields the expected dimensional structure of the measured construct in an equivalent manner across groups (see, e.g., Byrne & Watkins, 2003).[4] In both instances, the parameters of most interest are the factor covariances. Here again, these tests can be based on the analysis of covariance structures.

We have summarized the basic set of tests for cross-group measurement and structural equivalence based on SEM analyses. Although Meredith (1993) distinguished among three increasingly restrictive tests for equivalence, to date there is no concrete directive dictating the extensiveness of this testing procedure. Clearly, decisions regarding level of testing stringency will be determined a priori and tailored by both the focus and data of the study (see Widaman & Reise, 1997). (For detailed descriptions of these tests for multigroup equivalence, readers are referred to Horn & McArdle, 1992; Little, 1997; Widaman & Reise, 1997; for an annotated explanation and illustration of diverse models based on the LISREL, EQS, and AMOS programs to Byrne, 1998, 2006, 2009, respectively; Byrne, 2008; and for a review of this multigroup equivalence literature, to Vandenberg & Lance, 2000.)

Applications to Cross-Cultural Data

Several researchers have addressed the issue of equivalence in cross-cultural research (see, e.g., Byrne, 2003; Johnson, 1998; Leong, Okazaki, & Tak, 2003; Poortinga, 1989; Leung & Wong, 2003; van de Vijver & Leung, 1997, 2000) and all agree that it encompasses many complexities. (For a comprehensive elaboration of these complexities, readers are referred to Byrne, Oakland, Leong, van de Vijver, Hambleton, Cheung, & Bartram, 2009.) Although cross-cultural researchers have devised a variety of approaches in addressing issues of equivalence, most have not employed SEM in testing for such equivalence. Van de Vijver and Leung (1997, 2000) have urged researchers to embrace this methodological approach in testing for equivalence and have explicitly outlined how it can be implemented. Despite the work of these cross-cultural methodologists, together with relevant pedagogical papers on the topic (e.g., Byrne, 2003; Byrne & Campbell, 1999; Byrne & Stewart, 2006; Byrne & Watkins, 2003; Ryan, Chan, Ployhart, & Slade, 1999), there remains little evidence of SEM application to tests for equivalence in the cross-cultural literature. Furthermore, it is somewhat puzzling that, of the few SEM studies reporting findings based on tests for equivalence (one clear exception being Marsh et al., 2006), most have involved less than five cultures. Indeed, van de Vijver and Leung (2000) alluded to this low number in their summary of perceived methodological weaknesses in cross-cultural psychology.

Although there has been a modicum of studies that has applied Meredith's (1993) test for "strong" equivalence with cross-cultural data (e.g., Byrne, Stewart, Kennard, & Lee, 2007; Cooke et al., 2001; Little, 1997), in all cases, interest focused solely on the comparison of latent factor means across cultural groups, with the related tests for equivalence serving merely as necessary stepping stones to obtaining this information. Overall, there appears to be substantial agreement among many researchers that, unless there is specific interest in testing for latent mean group differences, tests for the equivalence of an assessment instrument, within the framework of SEM,[5] can realistically be limited to the factor loadings and relations among their underlying latent factors (see, e.g., Cooke et al., 2001; Marsh, 1994; Marsh et al., 2006).

TESTING FOR EQUIVALENCE IN LARGE-SCALE CROSS-CULTURAL STUDIES

Complexities Inherent in the Standard Approach to Testing for Equivalence

A number of complexities derive from the intrinsic shortcomings of using standard SEM multigroup procedures in testing for equivalence when applied to large-scale cross-cultural data. When tests for equivalence involve a small number of cultural

groups, researchers can first validate the factorial structure of the measuring instrument for each group separately before simultaneously testing for its invariance across the groups (for illustrated applications see Byrne, 1998, 2006, 2008, 2009). This process can sometimes lead to the specification of additional parameters for one group, but not for others. Although, technically speaking, it is not necessary that all parameters be tested for their equivalence across groups (Jöreskog, 1971) and in the case of partial measurement, that additionally specified parameters can be freely estimated for each group (Byrne et al., 1989), both options quickly become unwieldy when the number of cultural groups is large. This reality, then, leads to at least three important complications. First, in testing for the equivalence of hypothesized factorial structure, it is assumed that all samples derive from the same population. However, when the groups under study represent different countries, this assumption of a single parent population becomes increasingly invalid as the number and diversity of the groups grows. Second, this approach fails to alert researchers to the possible inappropriateness of the construct, as structurally and psychometrically postulated, for particular cultures. Third, standard procedures can deal with cross-cultural similarities but provide few strategies for dealing with nonequivalence. An inspection of modification indices may work well in dealing with a small number of groups, but these indices are often of limited value in data sets comprising a large number of groups as they focus on deviances per country, thereby preventing any integration of larger and psychologically more meaningful units, such as clusters of countries with similar religions or levels of affluence.

An Alternative Approach to Testing for Equivalence in Large-Scale Studies

A major limitation of all cross-cultural equivalence studies is the solitary focus on identification of problematic features in the *instrument* (i.e., evidence of item bias or nonequivalence). More appropriately, however, we contend that researchers also should consider the possibility that *particular countries* may be problematic and need to be excluded from the analyses. Alternatively, it may be that the multicultural sample consists of clusters of countries exhibiting both within-cluster homogeneity (i.e., the instrument exhibits equivalence across all members of the same cluster) and between-cluster heterogeneity (i.e., instrument nonequivalence across clusters), both of which demand that analyses be tailored accordingly.

As one alternative strategy that addresses this major limitation, we present a dual modal two-pronged approach to testing for equivalence in large-scale cross-cultural studies. Accordingly, we test first for the factorial validity of the measuring instrument and for the multigroup equivalence of this factorial structure (i.e., the configural SEM model) across 27 cultural groups (Module 1). Confronted with evidence of an ill-fitting configural model, we conduct a series of univariate, multivariate, and SEM analyses that focus on the extent to which particular items,

cultures, or a combination of both, contribute to the poor fit of the 27-culture configural model (Module 2).

METHOD

Samples and Procedures

Data used in our SEM equivalence example derive from a large project designed to measure family functioning across 30 cultures (Georgas, Berry, van de Vijver, Kagitcibasi, & Poortinga, 2006). Selection of countries in the project focused on representation of the major geographical and cultural regions of the world so as to maximize eco-cultural variation in known family-related context variables such as economic factors and religion (Georgas et al., 2006).Thus, countries were selected from north, central, and south America; north, east, and south Europe; north, central, and south Africa; the Middle East; west and east Asia; and Oceania. Our interest in the present study lies with responses to the Family Values (FV) Scale (Georgas, 1999) for 5,482 university students drawn from 27 of these 30 countries.[6] Data comprised 2,070 males and 3,160 females[7] ; ages ranged from 15 through 38 years, with the largest proportion ($n = 4{,}861$) falling between the ages of 18 and 26. Country data on sample size, gender composition, and mean age are reported in Table 1. Although sample sizes of some cultural groups were undisputedly small, there is some evidence that when the number of groups and total sample size are large, as is the case here, parameter estimates can remain adequately stable (see Cheung & Au, 2005).

The FV Scale was administered in university classroom settings and response data collected by the research team trained in each country. All members of each team were indigenous to their home country. Based on the SPSS Missing Value Analysis module, the relatively few missing values in the data were replaced by regression-based estimates to which an error component was added. Thus, all FV Scale item scores used in the present study were complete.

Instrumentation

Building on the work of Georgas (1989), the FV Scale is designed to measure the influence of modernization and urbanization (i.e., acculturation) on family values, as well as on the attitudes and behaviors of urban and rural Greeks, the expectation being that more urbanized individuals would hold more individualistic family values, whereas rural individuals would maintain more collectivistic (and traditional) family values. The version used in the current example measures only two factors: hierarchical roles of father and mother and relationships within the family and with kin.

TABLE 1
Sample Descriptives and Multivariate Non-normality across Cultural Groups

Country	Sample Size [a]	Females	Males	Mean Age	Multivariate Kurtosis (Mardia's Normalized Estimate)[b]
Algeria	107	66	41	20.97	53.56
Brazil	159	107	52	21.73	28.65
Bulgaria	195	117	78	21.64	27.22
Canada	215	159	56	19.24	21.46
Chile	207	104	103	21.57	13.76
Cyprus	132	114	18	20.49	16.46
France	97	86	11	21.19	8.27
Georgia	200	116	84	20.17	53.42
Germany	153	106	40	22.43	7.76
Ghana	70	16	54	27.16	15.35
Greece	350	243	107	21.34	29.47
Hong Kong	423	218	205	19.00	34.85
India	220	98	121	22.02	41.76
Indonesia	239	—	—	—	58.48
Iran	189	130	59	21.28	24.61
Japan	185	97	88	19.52	18.38
Mexico	227	124	102	22.73	30.87
Nigeria	337	137	197	23.93	138.39
Pakistan	450	238	212	19.82	156.95
Saudi Arabia	198	59	139	22.22	86.29
South Korea	199	120	79	20.85	27.52
Spain	111	85	26	19.09	8.68
The Netherlands	165	128	37	20.20	11.71
Turkey	211	165	46	19.36	21.41
Ukraine	65	50	14	20.79	7.41
United Kingdom	115	83	32	22.26	14.04
USA	263	194	69	21.20	46.48

[a]Sum of females and males can be smaller than total sample size due to missing values. [b]Values > 5.00 are indicative of multivariate non-normality (Bentler, 2005).

The FV Scale is an 18-item measure having a 7-point Likert scale that ranges from 1 (*strongly disagree*) to 7 (*strongly agree*). Items were derived from an original 64-item pool and selected in such a way that the expected factors (hierarchy and family/kin relationships) would be well represented. Based on EFA findings that revealed near-zero loadings for 4 items (see van de Vijver, Mylonas, Pavlopoulos & Georgas, 2006)[8] we included only 14 of the 18 items in our application; abbreviated content pertinent to these items can be viewed in Table 4.

Working from the English version of the FV Scale and using the adaptation approach to test translation (Harkness, 2003), all items and instructions were translated by the research team of each country into the target language of that country. During the translation process there was frequent contact between the local

researcher and the principal investigator (Georgas) to discuss translation problems. The items and instructions were then back-translated into English. In an effort to assess the extent to which the two translations were equivalent in connotation, the research team subsequently compared the back-translated items with the original English items. Any items exhibiting nonequivalence of meaning were discussed and then rephrased in the target language until linguistic equivalence was deemed satisfactory.

Internal consistency coefficients were computed by factor for the total sample; Cronbach's coefficient alpha was .87 for the Hierarchy Scale and .80 for the Relationships Scale. Country-wise analyses showed a median alpha coefficient of .78 (IQR = .10) for the first scale and .74 (IQR = .11) for the second scale.

The Hypothesized Model

The CFA model of FV Scale structure is shown schematically in Figure 1. This model hypothesized a priori that, for each cultural group: (a) the FV Scale is most appropriately represented by a 2-factor structure comprising the constructs of Family Hierarchy and Family/Kin Relationships, (b) each observed variable (i.e., FV Scale item) has a nonzero loading on the factor it was designed to measure, and zero loadings on all other factors, (c) the two factors are correlated, and (d) measurement error terms are uncorrelated.

Statistical Analyses

All SEM and some univariate analyses were based on the EQS 6.1 program (Bentler, 2005); all others were based on SPSS, Version 15. One critically important assumption underlying SEM analyses is that the data are multivariate normal. If, indeed, they are not, it is imperative that analyses be based on an estimation procedure capable of addressing such non-normality and, in particular, the presence of multivariate kurtosis, which can seriously distort parameter estimates (West, Finch, & Curran, 1995). Robust maximum likelihood estimation is recommended as the appropriate approach in addressing this problem as it performs well across different levels of non-normality, model complexity, and sample size (Curran, West, & Finch, 1996). Given that pre-analysis of the present data revealed evidence of moderately high levels of multivariate kurtosis across combined cultural groups and substantial multivariate kurtosis for particular cultural groups (see Table 1), all analyses were based on the EQS robust statistics. Accordingly, parameter estimation was based on robust maximum likelihood procedures and model fit evaluations on the Satorra-Bentler χ^2 (S-Bχ^2; Satorra & Bentler, 1988) and related robust fit indices (to be described later). The S-Bχ^2 serves as a correction for the χ^2 statistic when distributional assumptions are violated and has been shown to be the most reliable test statistic for evaluating mean and covariance

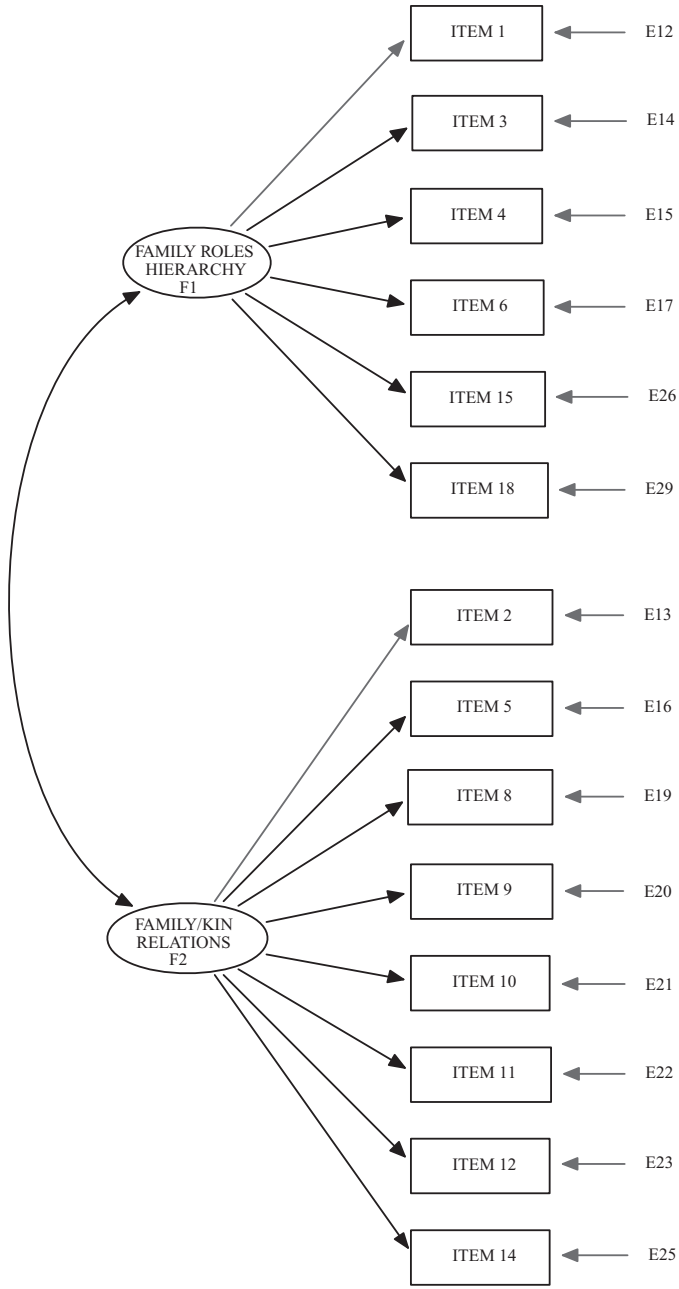


FIGURE 1
Hypothesized factorial structure of the Family Values Scale (Georgas, 1999).

structure models under various distributions and sample sizes (Hu, Bentler, & Kano, 1992; Curran et al., 1996). Given the focus of the present study on illustrating possible difficulties in the attainment of a well-fitting multicultural configural model, we determined analysis of only covariance structures to be appropriate for our purposes.

As noted earlier, our testing strategy is presented within the framework of two separate modules, the details of which are now outlined.

Module 1: Testing for the Validity and Equivalence of Factorial Structure

Step 1. We tested first for goodness-of-fit for the hypothesized two-factor structure based on the pooled variance/covariance matrix. That is, we ignored any between-group variability across culture in testing for the validity of factorial structure.

Step 2. We tested the applicability of the pooled-within structure to the 27 countries. In other words, we tested whether the structure that was found for all samples together in Step 1 would apply to each country separately.

Goodness-of-fit criteria. Evidence of model goodness-of-fit was based on triangulated findings from multiple indices as suggested by Hu and Bentler (1999); these include both the original and robust versions of the Comparative Fit Index (CFI; Bentler, 1990) and Root Mean Square Error of Approximation (RMSEA; Steiger, 1990), as well as the Standardized Root Mean Square Residual (SRMR). Given that computation of both the robust CFI (*CFI) and RMSEA (*RMSEA) are based on the S-Bχ^2 scaled statistic, their values are also corrected to take non-normality into account. The CFI ranges in value from zero to 1.00, with a value of .95 serving as the rule-of-thumb cutpoint of acceptable fit (see Hu & Bentler, 1999). The RMSEA takes into account the error of approximation in the population and is expressed per degree of freedom, thus making it sensitive to model complexity; values less than .05 indicate good fit, and values as high as .08 represent reasonable errors of approximation in the population. For completeness, we also include the 90% confidence interval provided for the RMSEA (Steiger, 1990). Finally, the SRMR is the average standardized residual value derived from fitting the hypothesized variance covariance matrix to that of the sample data. Its value ranges from zero to 1.00, with a value less than .08 being indicative of a well-fitting model (Hu & Bentler, 1999).

Module 2: Testing for Item versus Country Sources of Bias

Provided with findings of poor fit in testing for validity of the multigroup configural model, analyses then centered on pinpointing the source of the problem. Specifically, we used a two-pronged approach to identify the extent to which particular

items, cultures, or a combination of both contributed to the ill-fitting 27-country model as follows:

Step 1. We examined factor loadings, descriptive statistics, and inter-quartile ranges for each item and country in an effort to discern any visible pattern of misfit.

Step 2. A series of CFA models was tested in an effort to pinpoint model misfit at both the item and country levels. Given that the analyses in Module 2 build on one another, we consider their description in more detail in the Results section.

RESULTS

Module 1: Testing for the Validity and Equivalence of Factorial Structure

Findings for all SEM analyses conducted in Module 1 are summarized in Table 2. As shown in this table, fit of the two-factor structure to the pooled data (Model 1), yielded a fairly well-fitting model as indicated by both the ML (CFI = .936; RMSEA = .069; SRMR = .054) and Robust (*CFI = .939; *RMSEA = .057) estimates. These findings support the validity of the hypothesized 2-factor structure of the FV Scale.

Model 2, the configural model, tested simultaneously for the validity of the hypothesized 2-factor structure across each of the 27 cultures. As shown in Table 2, the ML and Robust CFI results were consistent in yielding estimates indicative of a very poor fit to the data (CFI = .852; *CFI = .837), albeit there appears to be some discrepancy between the ML and Robust RMSEA values. Whereas the ML RMSEA value of .089 for Model 2 is substantially higher than is the case for Model 1, this value discrepancy is much less pronounced for the robust estimates.

In cases where this standard SEM approach is used in testing across only two or three groups, the next logical step would be to identify the fixed parameters in the model contributing most to misfit. In EQS, this task is facilitated through implementation of the Lagrange Multiplier Test (LMTest). However, when the test involves many groups, identification of the appropriate misspecified parameters via examination of these modification indices is extremely difficult, if not impossible. In the current case, for example, despite limiting the possible misspecified parameters to only cross-loadings and error covariances, the LMTest results for Model 2 identified 325 statistically significant parameters that were possible sources of misfit! Furthermore, these tagged parameters varied widely across the 27 cultures. The two largest misfitting parameters represented an error covariance between Items 5 and 8 for Nigeria and a cross-loading of Item 1 on Factor 2 for Saudi Arabia (see abbreviated item content in Table 4). Incorporation of these two parameters into the model yielded a minimal drop in the S-B χ^2 value and virtually no change in either

TABLE 2
Goodness-of-fit Statistics for Module 1

	Maximum Likelihood Estimates						Robust Estimates			
Model Description	χ^2	df	CFI	RMSEA	90% CI	SRMR	S-Bχ^2	*CFI	*RMSEA	*90% CI
1. Test of 2-factor model based on pooled within-covariance matrix	2,066.69	76	.936	.069	.067; 072	.054	1,459.89	.939	.058	.055; 060
2. Test of configural model	5,361.78	2052	.852	.089	.086; .092	.088	3,854.42	.837	.066	.063; .069

CFI = Comparative Fit Index; RMSEA = Root mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; S-Bχ^2 = Satorra-Bentler Scaled Statistic.

TABLE 3
Configural Model: 18 Highest Lagrange Multiplier Chi-square Statistics

	Country	Parameter		χ^2	p
		Type	Description		
1	Saudi Arabia	Error Covariance	Items 8 and 5	137.832	.000
2	Nigeria	Factor Cross-loading	Item 1 on Factor 2	80.725	.000
3	Hong Kong	Error Covariance	Items 11 and 1	43.622	.000
4	Hong Kong	Error Covariance	Items 8 and 9	41.094	.000
5	Brazil	Error Covariance	Items 11 and 12	35.849	.000
6	Pakistan	Error Covariance	Items 15 and 4	36.462	.000
7	Pakistan	Factor Cross-loading	Item 18 on Factor 2	35.604	.000
8	Ghana	Error Covariance	Items 1 and 2	34.853	.000
9	Greece	Factor Cross-loading	Item 9 on Factor 1	34.732	.000
10	France	Factor Cross-loading	Item 11 on Factor 1	33.210	.000
11	Hong Kong	Error Covariance	Items 15 and 11	32.769	.000
12	South Korea	Error Covariance	Items 8 and 9	31.696	.000
13	Pakistan	Error Covariance	Items 2 and 3	28.763	.000
14	India	Factor Cross-loading	Item 2 on Factor 1	27.304	.000
15	Turkey	Factor Cross-loading	Item 11 on Factor 1	27.244	.000
16	Algeria	Error Covariance	Items 2 and 5	27.131	.000
17	Saudi Arabia	Error Covariance	Items 1 and 14	26.304	.000
18	Pakistan	Error Covariance	Items 4 and 6	26.155	.000

the *CFI or *RMSEA values. Of import here is that, of the 325 LMTest modification indices related to this model, 316 had probability values less than .05. In other words, indications are that it would likely require the addition of 316 parameters to the configural model before an adequate fit across all 27 cultures would be attained. To provide readers with a flavor of why the process of model improvement is so slow, a summary of the 18 highest LMTest χ^2 values is presented in Table 3.

In a summary of Module 1 analyses, we conclude that: (a) the FV Scale is well described by a two-factor structure as indicated by the goodness-of-fit results for Model 1; (b) no support is found for configural invariance. Thus, the question at this point is whether misfit of the configural model is due to the nonequivalence of particular items across countries, to the cultural influence of particular countries, or to a combination of both. In an effort to resolve this interesting dilemma, we turn now to Module 2.

Module 2: Testing for Item versus Country Sources of Bias

This module included a rigorous and diverse series of univariate, multivariate, and CFA analyses, all directed toward identifying a possible pattern of misspecification related to either particular items, countries, or both. We began by examining the

factor loadings for each country in terms of their means, standard deviation, skewness, kurtosis, and extreme values.

Next, to determine the number of countries having item scores falling outside the normal range, we compared these scores across the 27 countries based on four increasingly restrictive cutpoint criteria. First, we compared item raw scores and pinpointed countries having visibly apparent extreme values. Second, we singled out countries having item scores that fell beyond the upper and lower bounds of the 99% confidence interval around the total sample mean. Third, we reviewed box plots of item scores identifying countries for which these values were classified either as "outlier" or as "extreme" scores. Outliers were defined as having scores that were 1.5 box lengths below the first or above the third quartiles; 3 box lengths were used for identifying extremes. Finally, we tagged countries having item scores greater or lesser than the median taking into account an effect size of 2.5 (Cohen, 1988). Based on evidence consistent with at least one of these identification criteria, analyses detected 7 deviant countries—France, Ghana, Indonesia, Japan, Nigeria, Pakistan, and Saudi Arabia.

All remaining analyses were based on various respecified SEM tests of the configural CFA model. Based on goodness-of-fit criteria recommended by Hu and Bentler (1999), and conducted for each country separately, we tested first for how many and which parameters needed to be added to this multigroup 2-factor model in order to reach a CFI value of .95. Again, our interest here was in discerning a possible pattern of misspecification. As might be expected, results revealed a wide diversity in both the number and type of parameters across countries. Due to restrictions of space, these results are not presented here.

To determine which item(s) and country(ies) contributed most to the misfit of the model, we tested for goodness-of-fit by (a) deleting one item at a time with replacement and (b) one of the seven deviant countries at a time with replacement. We based this decision on the differences in CFI values, with a difference value (ΔCFI) equal to or less than 0.01 being indicative of a substantially "practical" improvement in fit (see Cheung & Rensvold, 2002). Given the known sample-size sensitivity of the chi-square statistic, together with substantial and increasing support for use of the ΔCFI value, we considered the latter to provide the more logical and reasonable measure of model improvement than the traditional chi-square difference ($\Delta\chi^2$) value. Analyses identified four items that if deleted from the model, exhibited CFI difference values greater than 0.01; these were Item 9 (ΔCFI = .021; Children should help with chores), Item 2 (ΔCFI = .016; Good relationships with relatives), Item 8 (ΔCFI = .015; Children should care for elderly parents), and Item 11 (ΔCFI = .014; Children should obey parents). Of the seven deviant countries, only Pakistan, with a ΔCFI value of .012 exceeded the recommended cutpoint of .01. Factor loadings in the pooled and Pakistani samples are presented in Table 4. As can be seen, the loadings in the Pakistani data are much lower than in the pooled sample. A closer inspection of the data indicated that all items had high means in

TABLE 4
Comparison of Standardized Estimates: Pakistan versus Other Countries[a]

Item	Abbreviated content	Other Countries	Pakistan	Difference
1	Father should be head of family	0.81	0.47	*0.35*
2	Should maintain good relations with relatives	0.54	0.52	0.02
3	Mother's place is at home	0.68	0.43	*0.26*
4	Mother should be go-between	0.68	0.42	*0.26*
5	Parents should teach children proper behavior	0.53	0.43	0.10
6	Father should handle the money	0.80	0.46	*0.33*
8	Children should take care of elderly parents	0.58	0.48	0.10
9	Children should help with chores	0.43	0.50	–0.07
10	Problems should be resolved within family	0.58	0.64	–0.07
11	Children should obey parents	0.72	0.58	0.14
12	Should honor and protect family's reputation	0.76	0.42	*0.33*
14	Children should respect grandparents	0.63	0.52	0.11
15	Mother should accept father's decisions	0.77	0.60	0.17
18	Father should be breadwinner	0.81	0.54	*0.27*
	Factor correlation	0.62	0.72	

[a] Pooled solution. Differences larger than 0.25 are in italics.

the Pakistani sample; the lowest item mean (of the 7-point scale) was 5.02, with many items having values above 6.0. Moreover, there were strong relations between the factor loading and item means on both factors. We therefore concluded that the deviant position of Pakistan was due to methodological reasons, mainly ceiling effects and restriction of range, rather than to substantive reasons. It is unclear as to what extent these high values reflect acquiescence (which is known to be higher in less affluent and more collectivistic societies such as Pakistan; Harzing, 2006), strong endorsement of the traditional family values, or a combination of both.

Having identified four items that if deleted from the model would improve the fit to the data our next step was to determine which combination of these items, if deleted, would lead to the most improvement in model fit. Given that Item 9 contributed the most to misfit, we based our deletions on this item combined with variants of Items 2, 8, and 11. Results revealed the paired deletion of Items 9 and 2 to raise the CFI value to .892, and the combination of Items 9, 2, and 11 to the higher CFI value of .906. As a consequence, these latter three items were subsequently deleted, leaving a resulting configural model specifying a 2-factor structure based on 11 items.

At this point, we needed to determine how many and which parameters were needed to bring the model closer to goodness-of-fit having a CFI value of .95. Based on the LMTest results, we re-estimated a series of models in which we specified 13 additional parameters, all of which represented those identified as contributing most to misspecification of the model. However, given concerns of

parsimony and a resulting CFI value of only .932, we decided to forego this approach in our quest for a better fitting configural model. Thus, we backtracked to the first respecified model of this post hoc set of models, which included one error covariance (Items 8 and 5) for Nigeria, and one cross-loading (Item 1 on Factor 2) for Saudi Arabia. The LMTest statistics for these two parameters were markedly higher than all remaining values and thus argued for their inclusion in the model. These additions to the 11-factor model yielded a CFI value of .914.

Our final 11-item model included the deletion of one country (Pakistan) from the analyses. In addition to earlier findings of Pakistan as one of the 7 most deviant countries, as well as having the highest level of multivariate kurtosis, a review of the LMTest statistics revealed many of the misfitting parameters to be associated with this country. Thus, we deleted Pakistan from the analyses, which resulted in an improved CFI value of .925. Although we fell short of our goal in reaching a CFI value of .95, we believe that this final value of .925 is indicative of a very rigorously obtained and parsimonious CFI and therefore represents our final modified configural model.

Our analysis led to a reduction from 14 items to 11 items. It is important to address the question of whether this item reduction has altered the underlying concept. Embretson (1983) has coined the term *construct representation* to indicate whether the items constitute an adequate representation of the construct. The three items eliminated in our analyses involved maintaining good relations with parents (Item 2), taking care of elderly parents (Item 8), and obeying parents (Item 11). It is remarkable that all three items come from the second factor that dealt with family and kin relations. The three items that were removed do not define a specific subdomain of family and kin relations. Rather, the items refer to attitudes and practices that are known to differ considerably across cultures. Relations with relatives and obedience to parents are more important in interdependent cultures than in independent cultures (Kagitcibasi, 2007), while taking care of elderly and needy parents is much less common in western than in nonwestern countries (Ho, 1996). Thus, the items that were removed showed large cross-cultural differences, but did not involve a specific relation subdomain. Therefore, we conclude that neither factor was altered in content by the item reduction.

DISCUSSION

The overarching purpose of this article was to illustrate the extent to which use of the standard SEM approach to testing for equivalence can be problematic when applied to large-scale and widely diverse cultural groups. We fully acknowledge that there may be circumstances whereby such cross-cultural testing presents no difficulties; examples might include situations where (a) the countries of interest are all located within the same global region (i.e., Europe, Middle East, North

America, Asia etc.); (b) the measuring instrument has a well-established factorial structure supported by evidence of strong construct validity both within and external to its country of origin; or (c) the instrument of measurement represents an achievement, rather than an affective scale. Nonetheless, based on our own experiences, we believe that researchers are typically very likely to encounter findings of nonequivalence that necessarily require both explanation and resolution. These nonequivalencies certainly will occur when the factor structure itself is dissimilar across groups (i.e., item scores for a particular group are more appropriately represented by an alternate number of factors and/or pattern of factor loadings). However, they can also occur when the same factor structure best fits all groups under study. Nonequivalence in the latter case can arise for a variety of reasons, notably if the sources of nonequivalence have a global (rather than item-specific) influence on instrument scores (e.g., differential perception of item content or interpretation of Likert scale anchors, differential familiarity with item scale format, differential meaningfulness and/or relevance of the measured construct, etc.).

As noted previously, one common presumption in testing for the equivalence of a measuring instrument across cultural groups is that sources of possible bias rest solely with the items. In contrast, we contend that when the groups under study represent different countries, such bias may be driven by the extent to which respondents from a particular country have inculcated its social values and mores. Thus, we described an alternate approach to testing for equivalence that targeted cultures in addition to items as possible sources of bias.

In general, potential difficulties associated with tests for equivalence in large-scale cross-cultural studies can be explained by methodological and statistical limitations of both the procedures and data employed. We turn now to a brief review of these issues.

Methodological Issues

Three aspects of SEM procedures used in testing for equivalence are of import. First, because measuring instruments are often group-specific in the way they operate, it has been customary to establish a baseline model before testing for multigroup equivalence. These models represent the best-fitting, albeit most parsimonious model representing data for a particular group. Although typically, these baseline models are the same for each group, they need not be (see Bentler, 2005; Byrne et al., 1989). For example, it may be that the best-fitting model for one group includes an error covariance or a cross-loading, but not so for other groups under study. 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.

Once a well-fitting baseline model has been established for each group separately, these final models are then combined to form the multigroup model,

commonly termed the configural model. Although this technique typically works well when the number of groups is small, it is definitely not the case when the number of groups is large and diverse.[9] For example, in the case of the example data used in this article, the baseline models were found to vary considerably across groups as evidenced from the broad array of misspecified parameters noted in Table 3.

Second, given the somewhat impossible task of determining baseline models for a large number of groups, one can instead begin with the establishment of the configural model, the first procedural step in testing for equivalence. As such the same hypothesized factorial structure is specified for all groups simultaneously. However, once again, tests for equivalence must be based on a well-fitting configural model. The challenge here is to adequately establish a multigroup model that is sufficiently well-fitting. Indeed, our own attempts to attain a model fit capable of yielding a CFI of .95 again attests to the difficulty of this task.

Finally, common to all SEM programs is the practice of testing for the equality of constrained parameters by comparing two groups at a time. For example, given four groups, the program initially compares Group 1 with Group 2, then with Group 3 and then with Group 4. The researcher must subsequently respecify the input file such that on the next run, Group 2 is compared with Group 3 and then, with Group 4. The final respecification and testing of the input file compares Group 3 with Group 4. (For an example application of this procedure, see Byrne & Campbell, 1999.) Thus, it is easy to see that conducting a comparison of pairs across 27 countries (even though the program would structure the first set of comparisons between Country 1 and the remaining 26 countries) is rendered an exceedingly tedious, if not impossible task! As a result, given a large number of groups, it is common practice to limit the specification of equality constraints such that all groups are compared only to Group 1.

Statistical Issues

Given that SEM procedures derive from large-scale theory, their use incurs a strong assumption of multivariate normality. Indeed, it is now well known that in the analysis of covariance structures, multivariate kurtosis imposes the most damaging effects. In particular, the standard errors can be seriously attenuated, thereby leading to an inappropriate assessment of overall model fit. Thus, presented with evidence of non-normal data, the researcher needs to base analyses on an appropriate estimation procedure that can take this non-normality into account. In the present study, for example, we based analyses on the robust statistics provided in the EQS program. Given the item content of the FV Scale, not surprisingly, our example data revealed evidence of strong and widely fluctuating degrees of kurtosis. From a substantive perspective, it is interesting to note that such non-normality was most evident among Muslim countries.

A second statistical issue that can lead to problematic results is that of sample size. Once again, given the assumption of multivariate normality in SEM analyses, research has shown the need for sample sizes of at least approximately 200 (Boomsma & Hoogland, 2001). The smaller the sample size, the higher the variability and less likely the data are to be multivariate normal. As evidenced from our example data, sample sizes for some cultural groups were very small and may have contributed in a major way to our difficulties in trying to establish a well-fitting configural model. However, as noted earlier, it has been shown that when the data comprise many groups and the overall sample size is large, parameter estimates of even the small groups remains relatively stable (Cheung & Au, 2005). Nonetheless, attainment of statistically adequate sample sizes for all groups under study will go a long way in reducing this source of difficulty.

Implications

Our study has implications for practitioners in the field of family counseling as well as for family researchers. Our analysis has shown that family values have components that are widely shared across countries. For example, we found strong evidence that the same two factors, hierarchy and family/kin relations, constitute family values. It is fair to assume that these are essential components of family functioning in all cultures. This reasoning is in line with arguments by Schwartz (1992) according to which values emerge as ways to solve universal problems, such as coordination of actions in human groups. The universal relevance of hierarchy and family relations in family functioning is relevant for both practitioners and researchers. However, our study also shows that high levels of equivalence were not found. Comparisons of attitudes and behaviors by individuals from different cultures at face value are fraught with difficulties. For example, obedience is important in all cultures, but there are important differences in obedience demanded from children (more obedience is demanded in interdependent cultures). The practical implication is that attitudes and behaviors that are less abstract and more concrete are more difficult to compare; the more concrete the attitudes and behaviors, the more cultural knowledge is required to understand these.

CONCLUSION

To this point, we have discussed and demonstrated the difficulties that can occur in using the standard SEM approach to testing for equivalence in large-scale cross-cultural studies. The question now is which path of action can researchers take to avoid, or at least minimize these difficulties. Indeed, the labor-intensive search we followed in our attempt to identify sources of noninvariance and misspecification in the malfitting configural model, although systematic, was nonetheless

tedious and somewhat impractical under normal circumstances. At this time, development of more refined approaches to addressing these difficulties is underway. Nonetheless, we do offer one caveat that most certainly will reduce the problematic nature of these equivalence tests and, in addition, propose one possible solution by way of analytic approach.

Turning first to the caveat, we argue that the stronger the construct validity of the measuring instrument for which cross-group equivalence is sought, the less likely researchers are to face difficulties in determining an adequately specified configural model, thereby making subsequent tests for equivalence possible (see, e.g., Marsh et al., 2006). Ideally, construct validity of the selected instrument should reflect its replicated factorial validity within its country of origin, and, if possible, within at least some of the other countries comprising the comparative study. It may seem attractive to entirely rely on fit statistics in multigroup testing for drawing conclusions about equivalence. However, the mechanical use of fit statistics can easily lead to erroneous conclusions; knowledge of the cultures studied is also important in reaching conclusions.

From the perspective of overall study design, we propose that clustering the cultural groups in some meaningful manner would seem to be a reasonable approach to reducing difficulties where the factorial structure of the measuring instrument potentially may be different. We suggest three possible approaches that might be taken. First, test the validity of hypothesized structure for each cultural group separately and then cluster the groups according to the similarity of resulting structures. Second, identify important contextual variables and then cluster the cultural groups accordingly. Indeed, Cohen (2009) has argued that rather than continuing the common trend of comparing cultural groups geographically (e.g., Eastern versus Western), it makes more sense to think of culture as representing different forms of culture (e.g., religion, socioeconomic status, region within a country). In testing for the equivalence of a measuring instrument across culture, it may be worthwhile to cluster the groups according to affluence, religion, and/or global region. Of course, selected clusters derived from this approach should derive from appropriately conducted a priori statistical analyses. Finally, one might consider basing the clustering of groups on an established classification system such as the GLOBE Society Clusters proposed by House, Hanges, Javidan, Dorfman, and Gupta (2004).

In our experience, as well as that of others as noted earlier, testing for equivalence of a measuring instrument in large-scale cross-cultural studies can be fraught with difficulties. Thus, it was in this spirit that the present paper was written. We are hopeful that in discussing and illustrating these difficulties, we have served to make the path to a sound outcome less onerous for other researchers interested in testing for measurement and structural equivalence across multiple cultures.

NOTES

1. These inconsistencies stem from the fact that there is no baseline model for the test of equivalent variance-covariance matrices, thereby making it substantially more restrictive than is the case for tests of equivalence related to sets of model parameters; as a result, any number of inequalities may possibly exist across the groups under study (Byrne, 2006; Bentler, personal communication, October, 2005).
2. We use the term "uniqueness" in the factor analytic sense to indicate that portion of error variance arising from some characteristic considered specific (or unique) to a particular indicator (i.e. observed) variable.
3. Within the field of cross-cultural research, however, metric equivalence is not considered to be analogous to measurement equivalence.
4. Wells and Marwell (1976) noted more than 30 years ago that measurement and theory are inseparately wed. Thus, one tests either the validity of a theory (assuming accurate measurements) or tests the validity of the measuring instrument (assuming an accurate theory), but cannot validate both simultaneously.
5. We make this distinction because, in cross-cultural research, for example, it is common practice to use exploratory factor analytic (EFA) and DIF techniques to test for evidence of scale equivalence prior to conducting a path analysis, which typically again is based on the multiple regression approach rather than the SEM approach.
6. For reasons of technical complexities, the data for South Africa, Botswana, and Mongolia could not be used and thus, these countries were eliminated from all analyses.
7. Gender scores were missing for Germany ($n = 7$), India ($n = 1$), Mexico ($n = 1$), Nigeria ($n = 3$), Ukraine ($n = 1$), and Indonesia ($n = 239$).
8. These four items were:
   Item 7: Parents shouldn't get involved in the private lives of their married children.
   Item 13: Parents should help their children financially.
   Item 16: Children should work in order to help the family.
   Item 17: Parents shouldn't argue in front of the children.
9. One exception to this generalization can be found in Marsh et al. (2006) wherein the measuring instrument under test for multigroup equivalence can be considered the gold standard of exceptionally sound development and construct validation.

REFERENCES

Austin, J. T., & Calderón, R. F. (1996). Theoretical and technical contributions to structural equation modeling: An updated annotated bibliography. *Structural Equation Modeling*, *3*, 105–175.

Bentler, P. M. (1978). The interdependence of theory, methodology, and empirical data: Causal modeling as an approach to construct validation. In D. B. Kandel (Ed.), *Longitudinal research on drug use: Empirical findings and methodological issues* (pp. 267–302). New York: Wiley.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Bentler, P. M. (2005). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software (www.mvsoft.com).

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: A festschrift in honor of Karl Jöreskog* (pp. 139–168). Lincolnwood, IL: Scientific Software.

Byrne, B. M. (1988). The Self Description Questionnaire III: Testing for equivalent factorial validity across ability. *Educational and Psychological Measurement*, *48*, 397–406.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Byrne, B. M. (2003). Testing for equivalent self-concept measurement across culture: Issues, caveats, and application. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self research* (Vol. 1, pp. 291–314). Greenwich, CT: Information Age.

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. (2009). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge/Taylor & Francis.

Byrne, B.M., Baron, P., & Balev, J. (1998). The Beck Depression Inventory: A cross-validated test of factorial structure for Bulgarian adolescents. *Educational and Psychological Measurement*, *58*, 241–251.

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*, 557–576.

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training. *Training and Education in Professional Psychology*, *3*, 94–105.

Byrne, B. M., & Shavelson, R. J. (1986). On the structure of adolescent self-concept. *Journal of Educational Psychology*, *78*, 474–481.

Byrne, B. M., & Shavelson, R. J. (1987). Adolescent self concept: Testing the assumption of equivalent structure across gender. *American Educational Research Journal*, *24*, 365–385.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement equivalence. *Psychological Bulletin*, *105*, 456–466.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup equivalence of a second-order structure: A walk through the process. *Structural Equation Modeling*, *13*, 287–321.

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.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement equivalence revisited. *Journal of Cross-cultural Psychology*, *34*, 155–175.

Cheung, M. W. -L., & Au, K. (2005). Applications of multilevel structural equation modeling to cross-cultural research. *Structural Equation Modeling*, *12*, 598–619.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Cohen, A. B. (2009). Many forms of culture. *American Psychologist*, *64*, 194–204.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooke, D. J., Kosson, D. S., & Michie, C. (2001). Psychopathy and ethnicity: Structural, item, and test generalizability of the Psychopathy Checklist-Revised (PCL-R) in Caucasian and African American participants. *Psychological Assessment*, *13*, 531–542.

Curran, P. J., West. S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

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

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

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

Georgas, J. (1999). Family as a context variable in cross-cultural psychology. In J. Adamopoulos & Y. Kashima (Eds.), *Social psychology and cultural context* (pp. 163–175). Beverly Hills, CA: Sage.

Georgas, J., Berry, J. W., van de Vijver, F. J. R., Kagitcibasi, C., & Poortinga, Y. H. (2006). *Families across cultures: A 30-nation psychological study*. Cambridge, United Kingdom: Cambridge University Press.

Harkness, J. A. (2003) Questionnaire translation. In J. A. Harkness, F.J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). New York: Wiley.

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*, 243–266.

Hershberger, S. L. (2003). The growth of structural equation modelling: 1994–2001. *Structural Equation Modeling*, *10*, 35–46.

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

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement equivalence in aging research. *Experimental Aging Research*, *18*, 117–144.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Thousand Oaks, CA: Sage.

Hu, L-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Hu, L-T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351–362.

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national research. *Zuma Nachrichten Spezial 3*, 1–40.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

Kagitcibasi, C. (2007). *Family, self, and human development across cultures: Theory and applications* (2nd ed.). Mahwah, NJ: Erlbaum.

Leong, F. T. L., Okazaki, S., & Tak, J. (2003). Assessment of depression and anxiety in East Asia. *Psychological Assessment*, *15*, 290–305.

Leung, P. W. L., & Wong, M. M. T. (2003). Measures of child and adolescent psychopathology in Asia. *Psychological Assessment*, *15*, 268–279.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76.

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 121–147). Mahwah, NJ: Erlbaum.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial equivalence: A multifaceted approach. *Structural Equation Modeling*, *1*, 5–34.

Marsh, H. W. (2007). Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (3rd ed., pp. 774–798). New York: Wiley.

Marsh, H. W., Hau, K-T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, *6*, 311–160.

Meredith, W. (1993). Measurement equivalence, factor analysis, and factorial equivalence. *Psychometrika*, *58*, 525–543.

Millsap, R. E., & Kwok, O-M. (2004). Evaluating the impact of partial factorial equivalence on selection in two populations. *Psychological Methods*, *9*, 93–115.

Poortinga, Y. H. (1989). Equivalence of cross cultural data: An overview of basic issues. *International Journal of Psychology*, *24*, 737–756.

Ryan, A. M., Chan, D., Ployhart, R. E., & Slade, L. A. (1999). Employee attitude surveys in a multinational organization considering language and culture in assessing measurement equivalence. *Personnel Psychology*, *52*, 37–58.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). Orlando, FL: Academic Press.

Selig, J. P., Card, N. A., & Little, T. D. (2008). Latent variable structural equation modelling in cross-cultural research: Multigroup and multilevel approaches. In F. J. R. van de Vijver, D. A., van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 93–119). Mahwah, NJ: Erlbaum.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling*, *3*, 93–104.

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

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-cultural Psychology*, *31*, 33–51.

van de Vijver, F. J. R., Mylonas, K., Pavlopoulos, V., & Georgas, J. (2006). Results: Cross-cultural analyses of the family. In J. Georgas et al. (Eds.), *Families across culture: A 30- nation psychological study* (pp. 126–185). Cambridge, UK: Cambridge Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement equivalence literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

Welkenhuysen-Gybels, J., van de Vijver, F. J. R., & Cambré, B. (2007). A comparison of methods for the evaluation of construct equivalence in a multigroup setting. In G. Loosveldt, M. Swyngedouw, & B. Cambré (Eds.), *Measuring meaningful data in social research* (pp. 357–371). Leuven, Belgium: Acco.

Wells, L. E., & Marwell, G. (1976). *Self-esteem: Its conceptualization and measurement*. Beverly Hills, CA: Sage.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement equivalence of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention* (pp. 281–324). Washington, DC: American Psychological Association.