Assessing Measurement Invariance in Cross-National Consumer Research

JAN-BENEDICT E. M. STEENKAMP
HANS BAUMGARTNER*

Assessing the applicability of frameworks developed in one country to other countries is an important step in establishing the generalizability of consumer behavior theories. In order for such comparisons to be meaningful, however, the instruments used to measure the theoretical constructs of interest have to exhibit adequate cross-national equivalence. We review the various forms of measurement invariance that have been proposed in the literature, organize them into a coherent conceptual framework that ties different requirements of measure equivalence to the goals of the research, and propose a practical, sequential testing procedure for assessing measurement invariance in cross-national consumer research. The approach is based on multisample confirmatory factor analysis and clarifies under what conditions meaningful comparisons of construct conceptualizations, construct means, and relationships between constructs are possible. An empirical application dealing with the single-factor construct of consumer ethnocentrism in Belgium, Great Britain, and Greece is provided to illustrate the procedure.

A fuller understanding of consumer behavior and further advancement of consumer research as an academic discipline requires that the validity of models of consumer behavior developed in one country (mostly the United States) be examined in other countries as well (Bagozzi 1994; Dholakia, Firat, and Bagozzi 1980). A key concern in extending theories and their associated constructs to other countries is whether the instruments designed to measure the relevant constructs are cross-nationally invariant (Hui and Triandis 1985). Measurement invariance refers to ''whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute'' (Horn and McArdle 1992, p. 117). If evidence supporting a measure's invariance is lacking, conclusions based on that scale are at best ambiguous and at worst erroneous. For example, cross-national differences in scale means might be due to true differences between countries on the underlying construct or due to systematic biases in the way people from different countries respond to certain items. Similarly, cross-national differences in relationships between scale scores could indicate real differences in structural relations between constructs or scaling artifacts, differences in scale reliability, or even nonequivalence of the constructs involved. Findings of no differences between countries are open to analogous alternative interpretations. As succinctly stated by Horn (1991, p. 119): ''Without evidence of measurement invariance, the conclusions of a study must be weak.''

Although a variety of techniques have been used to assess various aspects of measurement equivalence (cf. Hui and Triandis 1985), there is general agreement that the multigroup confirmatory factor analysis model (Jöreskog 1971) represents the most powerful and versatile approach to testing for cross-national measurement invariance. In some marketing studies, elements of this approach have been put to good use in assessing the cross-national comparability of consumer behavior and marketing measures (see, e.g., Durvasula et al. 1993; Kumar, Scheer, and Steenkamp 1995; Lastovicka 1982; Netemeyer, Durvasula, and Lichtenstein 1991; Steenkamp and Baumgartner 1995). In general, however, critical reviews of the literature have identified a lack of concern for measurement invariance in cross-national consumer research (Mullen 1995; Netemeyer et al. 1991; Parameswaran and Yaprak 1987). We believe that this unsatisfactory state of affairs is due to several factors, including

*Jan-Benedict E. M. Steenkamp is professor of marketing and marketing area coordinator, Catholic University of Leuven, Naamsestraat 69, 3000 Leuven, Belgium, and G f K Professor of International Marketing Research, Wageningen University, Hollandseweg 1, 6706 KN Wageningen, The Netherlands. Hans Baumgartner is associate professor of marketing, The Mary Jean and Frank P. Smeal College of Business Administration, The Pennsylvania State University, University Park, PA 16802. At the time this article was written, the second author was visiting professor, Catholic University of Leuven, Naamsestraat 69, 3000 Leuven, Belgium. The order of the authors' names is arbitrary; both contributed equally to this research. The authors thank the editor, associate editor, and three reviewers for their help during the review process.

(1) the bewildering array of types of measurement invariance that can be found in the literature, (2) the lack of an agreed-upon terminology to refer to the different kinds of measurement equivalence, (3) researchers' relative unfamiliarity with testing measurement models that incorporate the latent and observed variable means, (4) the considerable methodological complexities involved in testing for different kinds of measurement invariance, some of which have not even been discussed in the consumer behavior and marketing literatures, (5) researchers' uncertainty about the extent to which measures have to be equivalent in order for particular cross-national comparisons to be meaningful, and (6) the absence of clear guidelines as to how to ascertain whether or not a measure exhibits adequate cross-national invariance.

The purpose of this article is to address these problems by reviewing the various forms of measurement invariance that have been proposed in the literature, organizing them into a coherent conceptual framework that ties different requirements of measure equivalence to the goals of the research, and offering a practical, sequential testing procedure that should facilitate the assessment of measurement invariance in cross-national consumer research. Our framework is based on the confirmatory factor analysis model and applies to any situation in which data are collected in at least two countries and the same set of items is used to operationalize the construct(s) of interest. We will draw on the technical literature in such diverse fields as multivariate statistics, psychometrics, developmental and cross-cultural psychology, education, and marketing to provide consumer behavior researchers with a comprehensive, systematic, and integrative discussion of the relevant issues that have to be considered before one can conduct meaningful cross-national comparisons. No such framework is currently available in the consumer behavior literature, even though the topic is probably of considerable interest to researchers working in international marketing. We will also illustrate the proposed procedure with an empirical application so that researchers can see how the approach is used in practice and how they can apply it in their own research.

MEASUREMENT MODEL

The relationship between observed variables and hypothesized underlying constructs can be modeled using the confirmatory factor analysis (CFA) model (Bollen 1989). In the typical CFA model, the observed response x_i to an item i ($i = 1, \ldots, p$) is represented as a linear function of a latent construct ξ_j ($j = 1, \ldots, m$), an intercept τ_i, and a stochastic error term δ_i. Thus,

$$x_i = \tau_i + \lambda_{ij}\xi_j + \delta_i, \tag{1}$$

where λ_{ij} is the slope of the regression of x_i on ξ_j. The slope coefficient, or factor loading, defines the metric of measurement, as it shows the amount of change in x_i due to a unit change in ξ_j. The intercept τ_i, in contrast, indicates the expected value of x_i when $\xi_j = 0$ (cf. Sörbom 1974).

Assuming p items and m latent variables, and specifying the same factor structure for each country g ($g = 1, \ldots, G$), we get the following measurement model:

$$x^g = \tau^g + \Lambda^g\xi^g + \delta^g, \tag{2}$$

where x^g is a $p \times 1$ vector of observed variables (in country g), ξ^g is an $m \times 1$ vector of latent variables, δ^g is a $p \times 1$ vector of errors of measurement, τ^g is a $p \times 1$ vector of item intercepts, and Λ^g is a $p \times m$ matrix of factor loadings. It is assumed that $E(\delta^g) = 0$ and that $\mathrm{Cov}(\xi^g, \delta^{g\prime}) = 0$. Equation 2 shows that observed scores on the p items are a function of underlying factor scores, but that observed scores may not be comparable across countries because of different intercepts (τ_i^g) and scale metrics (λ_{ij}^g).

To identify the model, the latent constructs have to be assigned a scale in which they are measured. In cross-national research (more generally, multigroup analysis) this is done by setting the factor loading of one item per factor to one; the identification problem should not be solved by standardizing the variances of the ξ_j (Cudeck 1989). Items for which loadings are fixed at unity are referred to as marker (or reference) items. The same item(s) should be used as marker item(s) in each country.

Taking expectations of Equation 2 yields the following relation between the observed item means and the latent means:

$$\mu^g = \tau^g + \Lambda^g\kappa^g, \tag{3}$$

where μ^g is the $p \times 1$ vector of item means and κ^g is the $m \times 1$ vector of latent means (i.e., the means of ξ^g). Unfortunately, κ^g and τ^g cannot be identified simultaneously (Sörbom 1982). The addition of any constant c to κ_j^g can be compensated for by subtracting $c\lambda_{ij}$ from τ_i^g. In other words, there is no definite origin for the latent variables. To eliminate this indeterminacy, specific constraints on the parameters are necessary. One possibility is to fix the intercept of each latent variable's marker item to zero in each country. This equates the means of the latent variables to the means of their marker variables (i.e., $\mu_m^g = \kappa_m^g$, where m indicates that the item is a marker item). A second possibility is to fix the vector of latent means at zero in the reference country (i.e., $\kappa^r = 0$, where the superscript r indicates the reference country) and to constrain one intercept per factor to be invariant across countries (as explained below, this has to be done for an item whose factor loading is invariant across countries). The latent means in the other countries are then estimated relative to the latent means in the reference country. The two methods lead to an exactly identified model with respect to the item intercepts and latent construct means. If further restrictions are imposed on the model (e.g., all intercepts are specified to be invariant across countries), the intercepts and latent means are overidentified, and the fit of the means part of the model can be investigated.

In addition to the mean structure given by Equation 3, the covariance structure has to be specified. As usual, the variance-covariance matrix of x in country g, Σ^g, is given by:

$$\Sigma^g = \Lambda^g \Phi^g \Lambda^{g\prime} + \Theta^g, \quad \mathbf{(4)}$$

where Φ^g is the variance-covariance matrix of the latent variables in ξ^g and Θ^g is the variance-covariance matrix of δ^g (usually constrained to be a diagonal matrix).

The overall fit of the model is based on the discrepancy between the observed variance-covariance matrices S^g and the implied variance-covariance matrices $\hat{\Sigma}^g$, and the discrepancy between the observed vectors of means m^g and the implied vectors of means $\hat{\mu}^g$. See Sörbom (1974) for mathematical details.

FORMS OF MEASUREMENT INVARIANCE IN CROSS-NATIONAL RESEARCH

Levels of Invariance

Configural Invariance. The configural invariance approach is based on Thurstone's principle of simple structure (Horn, McArdle, and Mason 1983). In essence, this principle states that the pattern of salient (nonzero) and nonsalient (zero or near zero) loadings defines the structure of the measurement instrument. In terms of factorial invariance, the principle of simple structure implies that the items comprising the measurement instrument should exhibit the same configuration of salient and nonsalient factor loadings across different countries (cf. Horn and McArdle 1992).

Although, in principle, the nonsalient loadings need not be constrained to zero, this is commonly done in CFA. Configural invariance is supported if the specified model with zero loadings on nontarget factors (if any) fits the data well in all countries, all salient factor loadings are significantly and substantially different from zero, and the correlations between the factors (if any) are significantly below unity. The third requirement is necessary to show that there is discriminant validity between the (sub)factors comprising the construct under investigation. Note that no cross-country constraints are imposed on the magnitude of the *salient* factor loadings; only *nonsalient* loadings are (implicitly) specified to be equal across countries (i.e., zero).

Metric Invariance. Configural invariance does not indicate that people in different countries respond to the items in the same way, in the sense that obtained ratings can be meaningfully compared across countries. Metric invariance provides for a stronger test of invariance by introducing the concept of equal metrics or scale intervals across countries (Rock, Werts, and Flaugher 1978). If an item satisfies the requirement of metric invariance, difference scores on the item can be meaningfully compared across countries, and these observed item differences are indicative of similar cross-national differences in the underlying construct. Since the factor loadings carry the information about how changes in latent scores relate to changes in observed scores, metric invariance can be tested by constraining the loadings to be the same across countries:

$$\Lambda^1 = \Lambda^2 = \ldots = \Lambda^G. \quad \mathbf{(5)}$$

Scalar Invariance. Configural and metric invariance require only information about the covariation of the items in different countries. However, in many research settings it is also important to conduct mean comparisons across countries. In order for such comparisons to be meaningful, scalar invariance of the items is required (Meredith 1993).[1]

Scalar invariance implies that cross-national differences in the means of the observed items are due to differences in the means of the underlying construct(s). It addresses the question of whether there is consistency between cross-national differences in latent means and cross-national differences in observed means. Even if an item measures the latent variable with equivalent metrics in different countries (metric invariance), scores on that item can still be systematically upward or downward biased. Meredith (1995) refers to this as additive bias. Comparisons of country means based on such additively biased items are meaningless unless this bias is removed from the data (Meredith 1993). Scalar invariance is tested by imposing the following additional constraint on the model of metric invariance:

$$\tau^1 = \tau^2 = \ldots = \tau^G. \quad \mathbf{(6)}$$

Factor Covariance Invariance. Invariance may also be imposed on the factor covariances. This restriction is tested by imposing the following cross-national constraints:

$$\phi_{jk}^1 = \phi_{jk}^2 = \ldots = \phi_{jk}^G \quad (j = 1, \ldots, m;\ k = 1, \ldots, [j-1]). \quad \mathbf{(7)}$$

Factor Variance Invariance. Invariance of the factor variances is tested by the following:

$$\phi_{jj}^1 = \phi_{jj}^2 = \ldots = \phi_{jj}^G \quad (j = 1, \ldots, m). \quad \mathbf{(8)}$$

If both the factor variances and covariances are invariant, the correlations between the latent constructs are invariant across countries.

Error Variance Invariance. A final form of invariance that may be imposed on the measurement model is

[1]We use the term "scalar invariance" to refer to the equality of measurement intercepts. It should be noted that other terms are sometimes used in the literature and that some authors (e.g., Hui and Triandis 1985) use "scalar invariance" in a broader sense to refer to invariance of factor loadings and item intercepts. Since scalar invariance is only tested for items that are metrically invariant, there is little danger of inconsistency, although we think it is important to differentiate between the two forms of measurement invariance (Meredith 1993).

that the amount of measurement error is invariant across countries. This is tested by specifying that:

$$\Theta^1 = \Theta^2 = \ldots = \Theta^G. \quad \textbf{(9)}$$

If items are metrically invariant, and if the error variances and factor variances are cross-nationally invariant, the items are equally reliable across countries.

Full versus Partial Invariance

The tests described so far are omnibus tests of whether a given level of invariance is fully satisfied or not. In practical applications, full measurement invariance frequently does not hold, and the researcher should then ascertain whether there is at least partial measurement invariance.[2]

Lastovicka (1982) relaxed the assumption of full configural invariance. He showed that while different factor structures might emerge from an analysis of different cultural groups, a subset of the factors investigated could still be cross-nationally invariant. We will refer to this as partial configural invariance.[3] Lack of full configural invariance may be due to some of the items loading on different factors or some of the constructs failing to achieve discriminant validity in some countries.

Although Lastovicka (1982) discussed the situation in which a model exhibits only partial configural invariance, he did not consider the possibility of partial invariance of other model components. Full measurement invariance is particularly unlikely for the more stringent forms of invariance following configural invariance. For example, Horn (1991, p. 125) calls metric invariance ''a reasonable ideal . . . a condition to be striven for, not one expected to be fully realized,'' and Horn et al. (1983) consider it scientifically unrealistic.

As a compromise between full measurement invariance and complete lack of measurement invariance, Byrne, Shavelson, and Muthén (1989) proposed the concept of partial measurement invariance (see also Reise, Widaman, and Pugh 1993). Partial measurement invariance as used by these authors applies to factors that are configurally invariant (in a model of partial configural invariance, the subset of factors that are configurally invariant), and the problem first emerges when metric invariance is imposed on the model. In particular, Byrne et al. (1989) argued that *full* metric invariance was not necessary in order for further tests of invariance and substantive analyses, such as comparisons of factor means, to be meaningful, provided that at least one item (other than the one fixed at unity to define the scale of each latent construct) was metrically invariant. Partial metric invariance only requires cross-country invariance of the zero loadings and of some, but not necessarily all, of the salient loadings.

Ideally, a researcher will be able to rely on substantive considerations when deciding which loadings should not be constrained to be equal across countries. Unfortunately, such detailed knowledge is often unavailable in cross-national consumer research, and the researcher has to rely mainly on empirical criteria in respecifying a model. Modification indices (MIs) and expected parameter changes (EPCs) are particularly useful in this context. However, model respecifications should be conducted cautiously, and in line with other authors (e.g., Kaplan 1989; MacCallum, Roznowski, and Necowitz 1992), we recommend that invariance constraints be relaxed only when MIs are highly significant (both in absolute magnitude and in comparison with the majority of other MIs) and EPCs are substantial. In addition, researchers should evaluate the change in alternative indices of overall model fit, especially those that take into account model parsimony (Steiger 1990). In general, the number of model modifications should be kept low, and only those respecifications that correct for relatively severe problems of model fit should be introduced. This minimizes capitalization on chance and maximizes cross-validity of the model (MacCallum et al. 1992).

It might happen that an item that is used as a marker item (i.e., an item that serves to define the scale of a latent variable) turns out not to be metrically invariant across countries (Reise et al. 1993). If this is the case, another item that does exhibit metric invariance has to be selected to serve as the marker item.

If partial metric invariance is supported, partial scalar invariance can be tested. The intercepts of those items that are not metrically invariant across groups are left unconstrained across countries, while the intercepts of the other items are (initially) held invariant. It is possible that some items have invariant loadings but cross-nationally different intercepts. If the initial model of partial scalar invariance is rejected, MIs and EPCs can again be used to locate intercepts that are not cross-nationally invariant. The invariance constraints on these intercepts are relaxed in subsequent models.

It can be shown that at least one item besides the marker item has to have invariant factor loadings and invariant intercepts in order for cross-national comparisons of factor means to be meaningful.[4] If one could assume that the constraints needed to identify the model's mean structure (e.g., fixing the vector of latent means at zero in the reference country and constraining one intercept to be invariant across countries) were correct, invariant loadings and intercepts for only one item would be sufficient; but in order to test this assumption, metric and scalar invariance for at least one additional item is required. Ideally, a majority of factor loadings and intercepts will

[2]Even when an omnibus test (e.g., a chi-square difference test) indicates that a set of invariance constraints is reasonable, it is still advisable to check that none of the individual constraints is seriously violated.

[3]An early discussion of configural invariance testing, which is based on exploratory factor analysis and congruence coefficients, is provided by Anderson and Engledow (1977).

[4]A proof of this proposition is available from the authors.

be invariant across countries because in that case the latent means are estimated more reliably (i.e., they are based on many cross-nationally comparable items) and differences in latent means succinctly summarize the pattern of differences in observed means across countries.

Partial measurement invariance can also be investigated for the factor covariances, factor variances, and error variances. In testing for the equivalence of error variances, one would logically set free the invariance constraints on error variances of those items that were found not to have cross-nationally invariant factor loadings.

LINKING FORMS OF INVARIANCE TO THE GOALS OF THE STUDY

In this article, we distinguish between three goals of cross-national research: exploring the basic structure of the construct cross-nationally, making quantitative comparisons of means across countries, and examining structural relationships with other constructs cross-nationally. The minimum level of invariance required is different for each of these goals. That means that for some studies relatively weak forms of invariance are sufficient (although more stringent forms are always preferable because they further strengthen the conclusions), whereas for other studies more stringent forms of invariance are required.

If the purpose is to explore the basic meaning and structure of the construct cross-nationally, in order to establish whether a construct can be conceptualized in the same way across countries, the minimum requirement is that the same pattern of (zero and nonzero) factor loadings is found in the different countries (Horn et al. 1983). Although one may also require the loadings to be equal across countries, we argue that metric invariance is desirable but not strictly necessary for this purpose. Following Thurstone (1947), the most basic and fundamental conceptualization of a construct is the pattern of zero and nonzero loadings, not the particular magnitude of the nonzero loadings. If a loading is cross-nationally significant (statistically and practically), evidence is found that the item is related to the underlying construct in each country, although the specific magnitude of the effect of the construct may differ. In other words, a set of items has to be cross-nationally congeneric, not necessarily tau equivalent, in order to conclude that a construct can be conceptualized in the same way across countries (Labouvie 1980). However, the researcher should refrain from making quantitative comparisons until more stringent forms of invariance have been established.

As discussed earlier in the article, metric and scalar invariance for at least two items per construct (or per factor if the construct is multidimensional) is required if the goal is to conduct comparisons of means across countries (Byrne et al. 1989; Meredith 1993). If this is not the case, comparing scores cross-nationally is meaningless since the measurement scales are fundamentally different across countries. The presence or absence of mean differences might be due to either real differences (lack of) or additive bias and/or different scale metrics. Neither invariance of factor (co)variances nor invariance of error variances is necessary for comparing means (Horn and McArdle 1992; Meredith 1993).[5] Examples of recent studies in which cross-national comparisons of means were conducted without assessing either metric or scalar invariance include Childers and Rao (1992), Dahlstrom and Nygaard (1995), Dawar and Parker (1994), and Verhage, Yavas, and Green (1990). Durvasula et al. (1993) and Kumar et al. (1995) assessed metric invariance before conducting cross-national comparisons of means but did not test for scalar invariance.

Finally, when the purpose of the study is to relate the focal construct to other constructs in a nomological net, full or partial metric invariance has to be satisfied because the scale intervals of the latent constructs have to be comparable across countries. Scalar invariance is not required because no absolute comparisons of scale scores are conducted. If the researcher wants to compare standardized measures of association (correlation coefficients, standardized regression coefficients) across countries, factor variance invariance is required in addition to metric invariance (Pedhazur 1982). Examples of recent studies in which correlations or standardized regression coefficients were compared cross-nationally without presenting evidence of either metric invariance or invariance of factor variances include Lee and Green (1991) and Rhee, Uleman, and Lee (1996). Lack of error variance invariance does not create a problem as long as differences in measurement error are explicitly taken into account (which is the case in latent variable modeling). However, when measures of association between observed variables are compared across countries, the scale reliabilities should be about the same so that measurement artifacts do not bias the substantive conclusions.

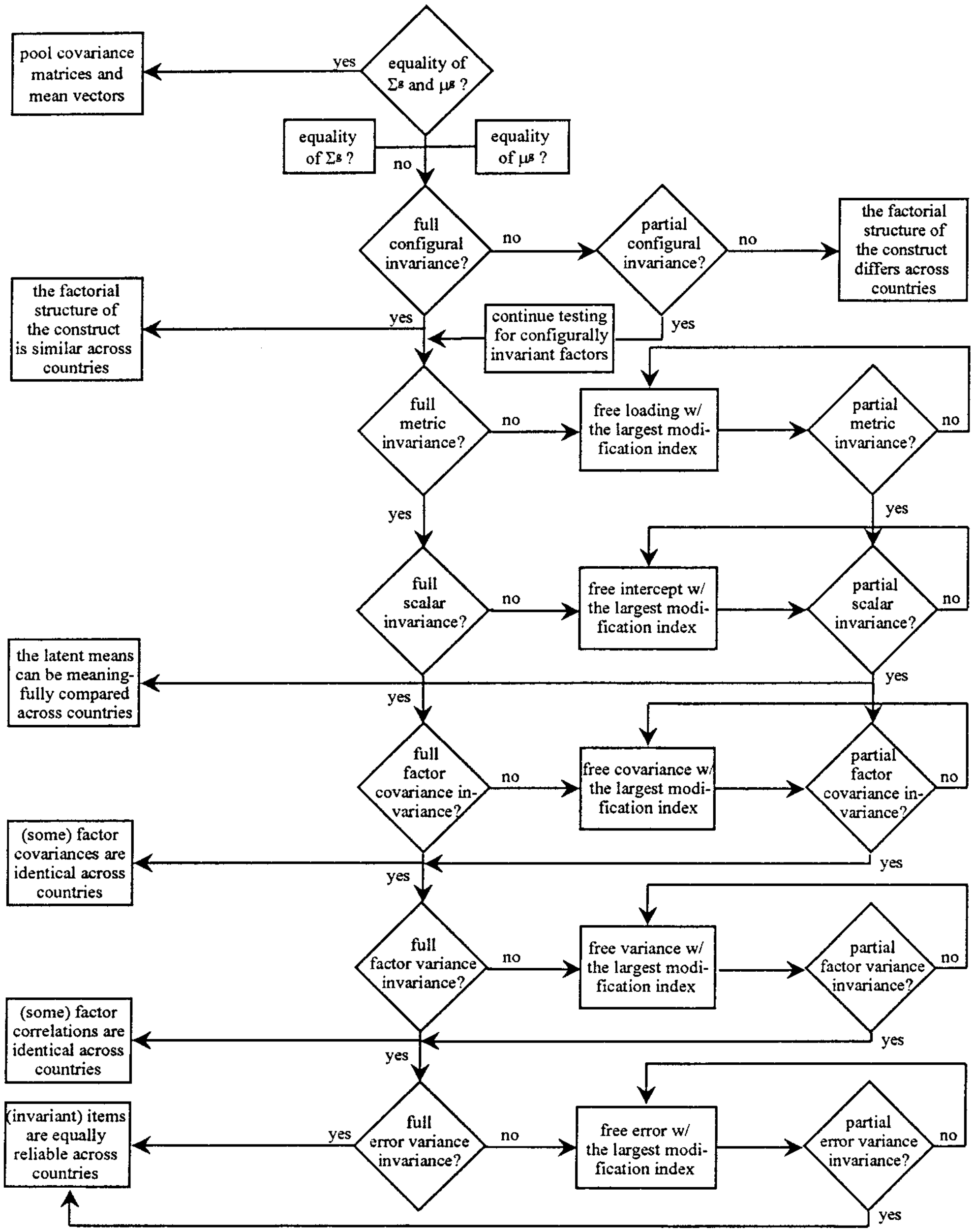
A PROPOSED PROCEDURE FOR TESTING MEASUREMENT INVARIANCE

Figure 1 contains a flowchart of the proposed procedure that researchers can use to assess the degree of cross-national invariance of their measurement instruments. The approach starts with a test of the equality of covariance matrices and mean vectors, both separately and jointly. It is not recommended that one test for the equality of moment matrices because even if

[5]Equal construct reliability is not necessary for mean comparisons (see, e.g., Rock et al. 1978). Construct reliability is affected by item loadings, error variances, and construct variances. However, by assumption, errors are random with an expectation of zero, so they should not affect the latent means. Furthermore, there is no conceptual or statistical reason why the construct variances should be equal across countries in order for comparisons of means to be meaningful. This leaves the factor loadings as the only remaining determinant of reliability, and their invariance is incorporated into our concept of measurement invariance.

FIGURE 1

PROPOSED PROCEDURE FOR ASSESSING MEASUREMENT INVARIANCE IN CROSS-NATIONAL CONSUMER RESEARCH



NOTE.—If the researcher is not interested in comparing means across countries, tests of scalar invariance can be omitted, and the analysis proceeds from assessing metric invariance to investigating factor covariance invariance.

the moment matrices are equal, there is no guarantee that the covariance matrices and mean vectors will be equal. These initial tests, which will probably show lack of invariance in most cross-national applications, provide useful information about whether the (co)variances or means are primarily responsible for the overall lack of invariance. In the unlikely case that the covariances and means are actually invariant across countries, the data can be pooled, and separate country analyses are unnecessary.

Whenever the covariance matrices and mean vectors are not invariant, configural invariance is examined first. The next step is to test for full metric invariance of those factors that exhibit configural invariance. If full metric invariance is satisfied, one can assess full scalar invariance; otherwise those loadings that are not metrically invariant should be left unconstrained across countries. Modification indices are the primary tool with which parameters are freed, although EPCs and change in overall fit indices should also be taken into account. If the model only exhibits partial metric invariance, tests of scalar invariance start with invariant intercepts for those items that are metrically invariant. Again, MIs are used to modify the model until a reasonable degree of partial scalar invariance has been achieved. At this point, one can conduct comparisons of the latent means across countries, provided at least two items per factor exhibit metric and scalar invariance.

Following MacCallum et al. (1994) and Marsh (1994), we propose the following sequence of tests for the remaining invariance constraints: factor covariance invariance, factor variance invariance, and error variance invariance. The covariances among the factors and the variances of the factors are typically of greater substantive interest than the error variances because they have a direct bearing on the magnitude of structural effects, even when corrected for measurement error. Furthermore, the covariances among the factors have important implications for the factor structure (e.g., in terms of discriminant validity), while the factor variances provide interesting information about the homogeneity of factor scores in the population. It is acknowledged, however, that the order of these tests is somewhat arbitrary (cf. Bollen 1989; Jöreskog 1971) and that it may depend on the purposes of the researcher. For example, if the focus is on measurement error and reliability, the reverse sequence might be more appropriate. Ultimately, the order of the three tests is not critical because in contrast to tests of configural, metric, and scalar invariance, the last three aspects of measurement invariance do not build on each other in the sense that one form of measurement equivalence has to be satisfied in order for subsequent tests to be meaningful.

It should be noted that, as in any other study that deals with cross-national consumer behavior, sample comparability is assumed across countries. If noncomparable samples are used, possible problems in measurement invariance are confounded with differences in the characteristics of the samples, which can lead to ambiguous interpretations. The two primary ways to achieve sample comparability are to draw nationally representative samples and to select matched samples on the basis of some set of characteristics of interest (see, e.g., Sekaran 1983).

With the exception of the model specifying equal covariances and equal means across countries, all models listed in Figure 1 can be placed in a hierarchical sequence of nested models so that systematic model comparison tests can be conducted. The standard way to compare the fit of competing models, provided they are nested, is the chi-square difference test (Jöreskog 1971). However, it has been observed that one should not rely exclusively on the chi-square difference test as it suffers from the same well-known problems as the chi-square test for evaluating overall model fit (see, e.g., Anderson and Gerbing 1988; Marsh and Grayson 1990). We therefore recommend using the following four alternative fit indices: the root mean square error of approximation (RMSEA); the consistent Akaike information criterion (CAIC), the Comparative Fit Index (CFI); and the Tucker-Lewis Index (TLI) or nonnormed fit index (NNFI; see Bagozzi and Baumgartner [1994] and Baumgartner and Homburg [1996] for details).[6] Smaller values of RMSEA and CAIC and larger values of CFI and TLI indicate better models. RMSEA, CAIC, and TLI seem particularly useful for purposes of model comparison because they take into account both goodness of fit and model parsimony by imposing a penalty on fitting additional parameters. In a recent simulation study, RMSEA, TLI, and especially AIC (on which CAIC improves) were found to be among the most effective indices in distinguishing between correctly and incorrectly specified models (Williams and Holahan 1994).

One disadvantage of the proposed testing procedure is that data-driven model modifications entail the danger of capitalizing on chance, which means that idiosyncrasies of a particular data set may lead to revisions of the originally hypothesized model that cannot be replicated with different data. Hence, cross-validation is strongly recommended (MacCallum et al. 1992; Steiger 1990).

EMPIRICAL ILLUSTRATION FOR CONSUMER ETHNOCENTRISM

We now present an example to illustrate the proposed procedure for assessing measurement invariance in cross-national consumer research. Our illustration deals with the measurement of consumer ethnocentrism in three European countries.

Method

Subjects. A pan-European market research agency collected nationwide data in three countries of the Euro-

[6] In LISREL, the fit of the null model used in computing TLI and CFI changes as different restrictions are imposed on the means part of the model (unless the means are exactly identified). This reduces the value of TLI and CFI for purposes of model comparison. Fortunately, the differences in the baseline model are generally not large, except in the case of testing for the equivalence of mean vectors and covariance matrices. Our recommendation is that incremental fit indices not be used in testing the hypotheses of equal means and of equal covariances and means across countries. Furthermore, if differences in the fit of the null model are large, TLI and CFI should be recalculated using a uniform null model (e.g., the model of uncorrelated observed variables). In our empirical application, the chi-square values for the null model varied very little across different model specifications. It should also be noted that in a recent unpublished article, Steiger (n.d.) argues that the multi-sample RMSEA values reported in LISREL (which were also used in this article) are incorrect. He maintains that in order to get the correct value, the LISREL figures have to be multiplied by a factor of $\sqrt{G}$. Future research will have to show whether this change affects the interpretation of RMSEA values in the multisample case.

pean Union, namely, Belgium, Great Britain, and Greece. Sample sizes were 990, 1,153, and 974, respectively. By random procedure, we split each country-sample in two and used the first half to estimate the models and the second half to cross-validate the results.

Measure. The construct that we focus on in this empirical illustration is consumer ethnocentrism. Consumer ethnocentrism can be defined as ''the beliefs held by consumers about the appropriateness, indeed morality, of purchasing foreign-made products'' (Shimp and Sharma 1987, p. 280). Shimp and Sharma (1987) developed and validated a 17-item scale to measure consumer ethnocentrism (called the CETSCALE), as well as a shorter, 10-item version (Shimp and Sharma 1987, n. 4; see also Netemeyer et al. [1991] for further validation evidence).

In our study, the 10-item version of the CETSCALE was used. Each item was rated on a five-point Likert scale. Mean scores and standard deviations on the CETSCALE for the three countries, using the raw scores, were as follows: Belgium, $\bar{X} = 28.70$, SD = 9.21; Great Britain, $\bar{X} = 30.29$, SD = 9.47; Greece, $\bar{X} = 37.84$, SD = 7.39.

Analysis. LISREL 8 (Jöreskog and Sörbom 1993) was used to analyze the covariances and means of the items.

Results

Calibration Data. As shown in Table 1, the test of equality of covariances and means yielded a chi-square value of 1,853.11 with 130 degrees of freedom ($p < .001$), an RMSEA of .0992, and a CAIC of 2,396.14. The statistics for the test of equality of covariances were: $\chi^2(110) = 1{,}137.60$ ($p < .001$), RMSEA = .0774, CAIC = 1,847.72, CFI = .922, TLI = .905; while the statistics for the test of equality of means were: $\chi^2(20) = 643.80$ ($p < .001$), RMSEA = .141, and CAIC = 2,105.81. It is apparent that the item means rather than the item covariances are the major determinant of the overall lack of invariance of the covariance matrices and mean vectors.

In line with Shimp and Sharma (1987), consumer ethnocentrism was conceptualized as a one-factor model. We fixed the scale and origin of the single latent variable by setting the loading of item 4 to one and its intercept to zero. The configural invariance model was estimated first. It is the baseline model against which the other models can be compared. The fit of the configural invariance model was satisfactory. Although the chi-square was significant ($\chi^2(105) = 936.09$, $p < .001$), the RMSEA of .0712 indicated an acceptable fit, and the two other practical fit indices were also above the commonly recommended .9 level (CFI = .937, TLI = .919). The CAIC for this model was 1,687.99. All factor loadings were highly significant in all countries, and 27 out of 30 (within-country) standardized factor loadings exceeded .6 (the minimum loading was .48). Thus, it can be concluded that the CETSCALE exhibited configural invariance across the three countries.

The hypothesis of full metric invariance was tested by constraining the matrix of factor loadings to be invariant across countries. From Table 1 it can be seen that there was a significant increase in chi-square between the model of configural and the model of full metric invariance ($\Delta\chi^2(18) = 142.36$, $p < .001$), although the fit did not decrease much in terms of the alternative fit indices. Examination of the MIs revealed that the significant increase in chi-square was due to a lack of invariance of four loadings that clearly stood out. The EPC statistics indicated that the factor loading of item 2 was much higher in Great Britain and the factor loadings of items 8 and 10 were much higher in Greece than in the other two countries, while the loading of item 6 was smaller in Greece than in the other countries. The MIs for these loadings were 45.21, 24.68, 19.89, and 44.05, respectively, in the model of full item-level metric invariance, although it should be noted that these values might change in the model modification process. Thus, full metric invariance was not supported.

To test for partial metric invariance, the constraints on these parameters were sequentially relaxed, starting with the loading that had the largest MI. The statistics for overall fit of the final model of partial metric invariance, after all four loadings were set free, are again reported in the upper half of Table 1. In terms of chi-square, the fit of this model is not significantly worse than the fit of the configural invariance model ($\Delta\chi^2(14) = 20.43$, $p > .10$); CFI is the same, while RMSEA, CAIC, and TLI have actually improved. Thus, partial metric invariance (with only four of 18 invariance constraints relaxed) is supported.

The next step was to impose scalar invariance on the model. However, given that only partial metric invariance was achieved, only the intercepts of the invariant factor loadings were constrained to be equal across countries. Scalar invariance for this model was not supported (see the upper half of Table 1). The increase in terms of chi-square was highly significant ($\Delta\chi^2(14) = 275.33$, $p < .001$), and the practical fit indices also showed a substantial deterioration in model fit. The MIs indicated that the intercept for item 2 (MI = 88.65) was not invariant across Greece and Belgium (note that this intercept was already unconstrained for Great Britain). Furthermore, the MIs suggested lack of invariance for the intercepts of items 1 (MI = 43.09) and 3 (MI = 55.77) in Greece. Successively relaxing these three constraints yielded a substantial and highly significant improvement in fit as compared to the full scalar invariance model: $\Delta\chi^2(3) = 207.49$, $p < .001$. Although the increase in chi-square relative to the partial metric invariance model (in which no constraints were imposed on the intercepts) is still significant ($\Delta\chi^2(11) = 67.84$, $p < .001$), model fit improved when considering RMSEA, CAIC, and TLI, while the decline in CFI was very small. There was no particular

TABLE 1

MODEL COMPARISONS FOR ETHNOCENTRISM DATA

	χ^2 value	*df*	RMSEA	CAIC	CFI	TLI
Calibration data:						
Equality of Σ^g and μ^g	1,853.11	130	.0992	2,396.14	n.a.	n.a.
Equality of Σ^g	1,137.60	110	.0774	1,847.72	.922	.905
Equality of μ^g	643.80	20	.1410	2,105.81	n.a.	n.a.
Configural invariance	936.09	105	.0712	1,687.99	.937	.919
Full metric invariance	1,078.45	123	.0706	1,679.96	.928	.921
Final partial metric invariance	956.52	119	.0672	1,591.45	.937	.928
Initial partial scalar invariance	1,231.85	133	.0728	1,749.82	.918	.917
Final partial scalar invariance	1,024.36	130	.0664	1,567.40	.932	.930
Full factor variance invariance	1,050.42	132	.0668	1,576.75	.931	.929
Final partial factor variance invariance	1,025.18	131	.0661	1,559.86	.932	.930
Initial partial error variance invariance	1,551.86	147	.0783	1,942.29	.894	.903
Final partial error variance invariance	1,098.95	142	.0657	1,541.73	.928	.931
Validation data:						
Configural invariance	952.42	105	.0721	1,703.79	.934	.915
Full metric invariance	1,058.15	123	.0700	1,659.25	.927	.920
Final partial metric invariance	983.49	119	.0684	1,617.98	.932	.923
Initial partial scalar invariance	1,257.36	133	.0738	1,774.97	.912	.913
Final partial scalar invariance	1,035.03	130	.0670	1,577.69	.929	.926
Full factor variance invariance	1,091.51	132	.0685	1,624.02	.925	.923
Final partial factor variance invariance	1,046.15	131	.0671	1,580.46	.928	.926
Initial partial error variance invariance	1,498.81	147	.0770	1,899.55	.894	.903
Final partial error variance invariance	1,152.70	142	.0677	1,595.17	.921	.925

NOTE.—n.a. means not applicable.

parameter that stood out on the basis of a highly significant MI. Finally, the model of partial scalar invariance was compared to the configural invariance model. The difference in terms of chi-square was significant ($\Delta\chi^2(25) = 88.27$, $p < .001$), but CFI decreased only an insubstantial .005, while RMSEA, CAIC, and TLI actually improved. Hence, it can be concluded that partial scalar invariance is supported.

Thus, valid comparisons of factor means can be conducted. The latent means for Belgium, Great Britain, and Greece were 2.639, 2.916, and 3.840, respectively (on a scale from 1 to 5). The model in which the three means were constrained to be invariant across countries yielded a significantly worse fit ($\Delta\chi^2(2) = 409.00$, $p < .0001$). Consumer ethnocentrism is significantly lower in Belgium than in Great Britain ($p < .0001$) and Greece ($p < .0001$), and it is significantly higher in Greece than in Great Britain ($p < .0001$).[7]

The hypothesis of invariant factor variances was rejected ($\Delta\chi^2(2) = 26.06$, $p < .001$). The MIs indicated that this was because of a difference in factor variance between Greece (MI = 22.21) and the other two countries. After removing the invariance constraint on the factor variance for Greece, the fit of the model was essentially the same as for the partial scalar invariance model: $\chi^2(131) = 1{,}025.18$ ($p < .001$), RMSEA = .0661, CAIC = 1,559.86, CFI = .932, TLI = .930. In a similar way, the initial model specifying partial invariance of error variances was rejected. The increase in chi-square was highly significant ($\Delta\chi^2(16) = 526.68$, $p < .001$), and the alternative fit indices also deteriorated substantially. After sequentially relaxing the invariance constraints on five error variances (one in Belgium and four in Greece; MIs in the initial model ranged from 46.73 to 105.74), the resulting model showed an adequate fit: $\chi^2(142) = 1098.95$ ($p < .001$), RMSEA = .0657, CAIC = 1,541.73, CFI = .928, TLI = .931. Note that this model achieves the best fit of all models considered so far, on the basis of all three fit indices that take into account model parsimony (RMSEA, CAIC, and TLI). The parameter estimates for this final model are shown in Table 2. The estimates of composite reliability are .939, .952, and .937 for Belgium, Great Britain, and Greece, respectively.

It can be seen in Table 2 that, on average, error and factor variances are smaller in Greece than in Belgium or Great Britain. This finding suggests that Greek consumers have firmer opinions about the morality of buying foreign-made products than consumers in the other two countries and that Greek consumers tend to agree more with each other in this respect. Indirect support for this conjecture is provided by a recent study conducted with European consumers (Commission of the European Communities 1995). A set of eight statements that differed in the extent to which they reflected national pride were presented to respondents, and they were asked to select the statement

[7]Valid comparison of factor means requires that the factor loading and the intercept of item 4 (i.e., the item used to define the scale and origin of the latent variable) be invariant across countries. The MIs indicated that this was indeed the case.

TABLE 2

ESTIMATION RESULTS FOR ETHNOCENTRISM DATA (CALIBRATION SAMPLE)

	Factor loadings			Error variances			Item intercepts		
Item	Belgium	Great Britain	Greece	Belgium	Great Britain	Greece	Belgium	Great Britain	Greece
ET1	.86	.86	.86	1.06	.53	.53	.94	.94	.59
	(33.62)	(33.62)	(33.62)	(15.22)	(22.01)	(22.01)	(11.91)	(11.91)	(5.82)
ET2	.71	1.04	.71	.71	.44	.71	1.53	.07	.86
	(21.74)	(31.16)	(21.74)	(21.60)	(15.19)	(21.60)	(15.96)	(.67)	(6.57)
ET3	.82	.82	.82	.60	.60	1.02	.43	.43	−.05
	(32.23)	(32.23)	(32.23)	(21.97)	(21.97)	(15.40)	(5.53)	(5.53)	(−.47)
ET4	1.00	1.00	1.00	.30	.30	.30	.00	.00	.00
	...	...	...	(24.41)	(24.41)	(24.41)	...	...	...
ET5	1.06	1.06	1.06	.33	.33	.14	−.26	−.26	−.26
	(56.07)	(56.07)	(56.07)	(19.62)	(19.62)	(12.60)	(−4.12)	(−4.12)	(−4.12)
ET6	1.09	1.09	.92	.32	.32	.14	−.08	−.08	.47
	(45.27)	(45.27)	(33.78)	(19.26)	(19.26)	(13.30)	(−1.04)	(−1.04)	(4.58)
ET7	1.02	1.02	1.02	.31	.31	.13	.01	.01	.01
	(55.95)	(55.95)	(55.95)	(19.70)	(19.70)	(12.57)	(.08)	(.08)	(.08)
ET8	.80	.80	.97	.54	.54	.20	1.04	1.04	.13
	(30.81)	(30.81)	(31.16)	(21.89)	(21.89)	(13.81)	(13.34)	(13.34)	(1.07)
ET9	.87	.87	.87	.56	.56	.19	.69	.69	.69
	(45.50)	(45.50)	(45.50)	(21.76)	(21.76)	(14.14)	(10.46)	(10.46)	(10.46)
ET10	.92	.92	1.07	.40	.40	.27	−.08	−.08	−.35
	(38.11)	(38.11)	(29.59)	(20.92)	(20.92)	(14.06)	(−1.07)	(−1.07)	(−2.53)

	Belgium	Great Britain	Greece
Variances	.95 (19.22)	.95 (19.22)	.62 (13.71)
Latent means	2.64 (55.93)	2.92 (66.76)	3.84 (95.16)

NOTE.—*t*-values are given in parentheses.

that came closest to their own opinion. Fully 56 percent of Greek respondents chose the strongest statement (''National pride is a duty for every citizen'') versus only 19 percent in Great Britain and 13 percent in Belgium. In the latter two countries, the opinions were more spread out across the statements.

Validation Data. Although the importance of cross-validation of modified models has been repeatedly stressed in the literature, it is rarely done in practice (MacCallum et al. 1992), and with the exception of Lastovicka (1982), who cross-validated his partial configural invariance model, we are not aware of cross-validation in partial invariance research. We reestimated the various models tested in the calibration data set using the validation sample.

The results are reported in the bottom half of Table 1. The fit measures in the calibration and validation samples are in close correspondence across all models tested, and the latent means are also very similar in both samples. As an even more stringent test of the stability of the model estimates, we used the parameter estimates from the final model of partial error variance invariance in the calibration sample on the validation sample. The fit indices for this model were as follows: $\chi^2(195) = 1{,}327.25$ ($p < .001$), RMSEA = .0612, CAIC = 1,327.25, CFI = .911, TLI = .939. Compared to the fit of the final model estimated on the validation sample, the difference in chi-square is significant ($\Delta\chi^2(53) = 174.55$, $p < .001$) but actually quite modest, given the large number of degrees of freedom involved (the ratio of chi-square over degrees of freedom actually declines by 16 percent). Moreover, the parsimony fit indices improve quite substantially, indicating that the decline in model fit is more than compensated for by the reduction in the number of parameters estimated.

DISCUSSION

It is important for scientific inference to have evidence of measurement equivalence. Such evidence is often not presented in cross-national research in the behavioral sciences (Horn and McArdle 1992; Hui and Triandis 1985). Lack of evidence of measurement invariance equivocates conclusions and casts doubt on the theory (Horn and McArdle 1992). In this article, we attempted to promote greater concern with measurement equivalence by providing an integrative overview of the various facets of cross-national measurement invariance and by describing how measure equivalence can be tested within the confirmatory factor analysis framework. A sequential procedure was proposed that can assist researchers in invariance testing. Special attention was given to partial measurement invariance as this is not a widely known concept among consumer behavior researchers, even though it is likely to be the typical case in many research situations. The proposed testing procedure allows researchers to examine partial measurement invariance in a systematic way, and if measurement instruments are at least partially invariant, valid cross-national comparisons can be conducted even when the ideal of full invariance is not realized. We illustrated the sequential testing procedure with data on consumer ethnocentrism and cross-validated the model of partial measurement invariance.

In the present application, the conclusions derived from the comparison of latent means were substantively the same as those obtained by comparing raw means. However, this need not always be the case. We have encountered situations in which the findings based on raw scores were quite misleading. One example concerned data for the construct of attitude toward advertising, using the five-item Gaski and Etzel (1986) scale. The data were collected in large nationally representative samples of Dutch, Danish, French, and Portuguese consumers. Items were scored on a five-point disagree/agree scale. One-way ANOVA on the raw scores revealed that the country means were not significantly different at $p < .05$. We also analyzed the data using the procedure described in this article, and very different conclusions were obtained. The latent means were significantly different from each other ($p < .001$), with Portugal and the Netherlands having significantly lower latent means (i.e., less positive attitudes toward advertising) than Denmark and France. France was not significantly different from Denmark, but Dutch consumers had a significantly lower attitude toward advertising than did Portuguese consumers. The substantive differences between latent means and raw means were due to strong additive bias in some of the items. Failure to take this response bias into account would lead to the erroneous conclusion that there are no differences between the four countries on attitude toward advertising.

Even when the raw means happen to lead to the same conclusions as the latent means, there is no excuse for not performing the necessary invariance tests. The crucial issue is not whether comparing or correlating raw scores leads to the same substantive results as does comparing or correlating factors scores (based on properly constrained models imposing the necessary invariance restrictions), but whether measurement invariance was investigated at all and whether the measurement instruments are sufficiently invariant to make cross-national comparisons meaningful. There are many examples in the literature in which the required tests of invariance were not performed, and the results of these studies must necessarily be ambiguous.

Several issues for future research can be identified. More research is needed on procedures that can be employed in the data-collection process to increase cross-national invariance in responses, including providing examples, describing and anchoring scales, and using the same data-collection procedures (see, e.g., Sekaran 1983). Another question is whether cultural factors can be identified that systematically contribute to cross-national differences in scale use. If that were the case, measurement invariance could be predicted a priori, and lack of invariance would be grounded in theory rather than diagnosed in the data. Another issue is the effect of such factors as number of scale items and sample size on the cross-validity of the sequential testing procedure described in this article and on the resulting parameter estimates.

A final issue is the comparison of competing models specifying various forms of measurement invariance and the appropriate determination of parameters that are allowed to differ across countries. We endorse the recommendations of Anderson and Gerbing (1988) to base model comparison on multiple fit indices, but this may bring an element of arbitrariness to the testing of alternative forms of invariance. In a similar way, all procedures for testing partial measurement invariance, including the use of MIs and EPCs, have their limitations. What is the right cutoff to stop relaxing invariance constraints, and is it dependent on other factors? In our experience, concentrating on a small number of large MIs and EPCs that clearly stand out has proven to be a robust and promising heuristic that can help in model respecification (see also Kaplan 1989; MacCallum et al. 1992), but more research is clearly necessary. Although important issues remain for future research, we hope that consumer behavior researchers will find the proposed framework to be a useful guide in assessing the cross-national invariance of their measurement instruments and that greater concern for measure equivalence will improve the methodological quality of cross-national consumer research.

[*Received August 1996. Revised August 1997. Robert E. Burnkrant served as editor and Richard P. Bagozzi served as associate editor for this article.*]

REFERENCES

Anderson, James C. and David W. Gerbing (1988), ''Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach,'' *Psychological Bulletin,* 103 (May), 411–423.

Anderson, Ronald and Jack Engledow (1977), ''A Factor Analytic Comparison of U.S. and German Information Seekers,'' *Journal of Consumer Research,* 3 (March), 185–196.

Bagozzi, Richard P. (1994), ''ACR Fellow Speech,'' in *Advances in Consumer Research,* Vol. 21, ed. Chris T. Allen and Deborah R. John, Provo, UT: Association for Consumer Research, 8–11.

——— and Hans Baumgartner (1994), ''The Evaluation of Structural Equation Models and Hypothesis Testing,'' in *Principles of Marketing Research,* ed. Richard P. Bagozzi, Cambridge, MA: Blackwell, 386–422.

Baumgartner, Hans and Christian Homburg (1996), ''Applications of Structural Equation Modeling in Marketing and Consumer Research: A Review,'' *International Journal of Research in Marketing,* 13 (April), 139–161.

Bollen, Kenneth A. (1989), *Structural Equations with Latent Variables,* New York: Wiley.
Byrne, Barbara M., Richard J. Shavelson, and Bengt Muthén (1989), ''Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance,'' *Psychological Bulletin,* 105 (May), 456–466.
Childers, Terry L. and Akshay R. Rao (1992), ''The Influence of Familial and Peer-Based Reference Group on Consumer Decisions,'' *Journal of Consumer Research,* 19 (September), 198–211.
Commission of the European Communities (1995), *Eurobarometer,* no. 42, Brussels: Commission of the European Communities.
Cudeck, Robert (1989), ''Analysis of Correlation Matrices Using Covariance Structure Models,'' *Psychological Bulletin,* 105 (March), 317–327.
Dahlstrom, Robert and Arne Nygaard (1995), ''An Exploratory Investigation of Interpersonal Trust in New and Mature Market Economies,'' *Journal of Retailing,* 71 (Winter), 339–361.
Dawar, Niraj and Philip Parker (1994), ''Marketing Universals: Consumers' Use of Brand Name, Price, Physical Appearance, and Retailer Reputation as Signals of Product Quality,'' *Journal of Marketing,* 58 (April), 81–95.
Dholakia, Nikhilesh, A. Fuat Firat, and Richard P. Bagozzi (1980), ''The De-Americanization of Marketing Thought,'' in *Theoretical Developments in Marketing,* ed. Charles W. Lamb and Patrick M. Dunne, Chicago: American Marketing Association, 25–29.
Durvasula, Srinivas, J. Craig Andrews, Steven Lysonski, and Richard G. Netemeyer (1993), ''Assessing the Cross-National Applicability of Consumer Behavior Models: A Model of Attitude toward Advertising in General,'' *Journal of Consumer Research,* 19 (March), 626–636.
Gaski, John F. and Michael J. Etzel (1986), ''The Index of Consumer Sentiment toward Marketing,'' *Journal of Marketing,* 56 (July), 71–81.
Horn, John L. (1991), ''Comments on 'Issues in Factorial Invariance,''' in *Best Methods for the Analysis of Change,* ed. Linda M. Collins and John L. Horn, Washington, DC: American Psychological Association, 114–125.
——— and J. Jack McArdle (1992), ''A Practical and Theoretical Guide to Measurement Invariance in Aging Research,'' *Experimental Aging Research,* 18 (Fall-Winter), 117–144.
———, Jack McArdle, and Ralph Mason (1983), ''When Is Invariance Not Invariant: A Practical Scientist's Look at the Ethereal Concept of Factor Invariance,'' *The Southern Psychologist,* 1 (Summer-Fall), 179–188.
Hui, C. Harry and Harry C. Triandis (1985), ''Measurement in Cross-Cultural Psychology: A Review and Comparison of Strategies,'' *Journal of Cross-Cultural Psychology,* 16 (June), 131–152.
Jöreskog, Karl G. (1971), ''Simultaneous Factor Analysis in Several Populations,'' *Psychometrika,* 36 (December), 409–426.
——— and Dag Sörbom (1993), *LISREL 8: User's Reference Guide,* Chicago: Scientific Software International.
Kaplan, David (1989), ''Model Modification in Covariance Structure Analysis: Application of the Expected Parameter Change Statistic,'' *Multivariate Behavioral Research,* 24 (July), 285–305.
Kumar, Nirmalya, Lisa K. Scheer, and Jan-Benedict E. M. Steenkamp (1995), ''The Effects of Supplier Fairness on Vulnerable Resellers,'' *Journal of Marketing Research,* 32 (February), 54–65.
Labouvie, Erich W. (1980), ''Identity versus Equivalence of Psychological Measures and Constructs,'' in *Aging in the 1980s: Psychological Issues,* ed. Leonard W. Poon, Washington, DC: American Psychological Association, 493–502.
Lastovicka, John L. (1982), ''On the Validation of Lifestyle Traits: A Review and Illustration,'' *Journal of Marketing Research,* 19 (February), 126–138.
Lee, Chol and Robert T. Green (1991), ''Cross-Cultural Examination of the Fishbein Behavioral Intention Model,'' *Journal of International Business Studies,* 21 (2), 289–305.
MacCallum, Robert C., Mary Roznowski, Corinne M. Mar, and Janet V. Reith (1994), ''Alternative Strategies for Cross-Validation of Covariance Structure Models,'' *Multivariate Behavioral Research,* 29 (January), 1–32.
———, Mary Roznowski, and Lawrence B. Necowitz (1992), ''Model Modifications in Covariance Structure Analysis: The Problem of Capitalization on Chance,'' *Psychological Bulletin,* 111 (May), 490–504.
Marsh, Herbert W. (1994), ''Confirmatory Factor Analysis Models of Factorial Invariance: A Multifaceted Approach,'' *Structural Equation Modeling,* 1 (1), 5–34.
——— and David Grayson (1990), ''Public/Catholic Differences in the High School and Beyond Data: A Multigroup Structural Equation Modeling Approach to Testing Mean Differences,'' *Journal of Educational Statistics,* 15 (Fall), 199–235.
Meredith, William (1993), ''Measurement Invariance, Factor Analysis, and Factorial Invariance,'' *Psychometrika,* 58 (December), 525–543.
——— (1995), ''Two Wrongs May Not Make a Right,'' *Multivariate Behavioral Research,* 30 (January), 89–94.
Mullen, Michael R. (1995), ''Diagnosing Measurement Equivalence in Cross-National Research,'' *Journal of International Business Studies,* 26 (3), 573–596.
Netemeyer, Richard G., Srinivas Durvasula, and Donald Lichtenstein (1991), ''A Cross-National Assessment of the Reliability and Validity of the CETSCALE,'' *Journal of Marketing Research,* 28 (August), 320–327.
Parameswaran, Ravi and Attila Yaprak (1987), ''A Cross-National Comparison of Consumer Research Measures,'' *Journal of International Business Studies,* 18 (Spring), 61–73.
Pedhazur, Elazar J. (1982), *Multiple Regression in Behavioral Research,* 2d ed., New York: Holt, Rinehart & Winston.
Reise, Steven P., Keith F. Widaman, and Robin H. Pugh (1993), ''Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance,'' *Psychological Bulletin,* 114 (November), 552–566.
Rhee, Eun, James S. Uleman, and Hoon Koo Lee (1996), ''Variations in Collectivism and Individualism by Ingroup and Culture: Confirmatory Factor Analyses,'' *Journal of Personality and Social Psychology,* 71 (November), 1037–1054.
Rock, Donald A., Charles E. Werts, and Ronald L. Flaugher (1978), ''The Use of Analysis of Covariance Structures for Comparing the Psychometric Properties of Multiple Variables across Populations,'' *Multivariate Behavioral Research,* 13 (October), 403–418.

Sekaran, Uma (1983), ''Methodological and Theoretical Issues and Advancements in Cross-Cultural Research,'' *Journal of International Business Studies,* 14 (Fall), 61–73.

Shimp, Terence A. and Subhash Sharma (1987), ''Consumer Ethnocentrism: Construction and Validation of the CETSCALE,'' *Journal of Marketing Research,* 24 (August), 280–289.

Sörbom, Dag (1974), ''A General Method for Studying Differences in Factor Means and Factor Structure between Groups,'' *British Journal of Mathematical and Statistical Psychology,* 27, 229–239.

——— (1982), ''Structural Equation Models with Structured Means,'' in *Systems under Indirect Observations,* pt. 1, ed. Karl G. Jöreskog and Herman Wold, Amsterdam: North-Holland, 183–195.

Steenkamp, Jan-Benedict E. M. and Hans Baumgartner (1995), ''Development and Cross-National Validation of a Short-Form of CSI as a Measure of Optimum Stimulation Level,'' *International Journal of Research in Marketing,* 12 (July), 97–104.

Steiger, James H. (1990), ''Structural Model Evaluation and Modification: An Interval Estimation Approach,'' *Multivariate Behavioral Research,* 25 (April), 173–180.

——— (n.d.), ''A Note on Multiple Sample Extensions of the RMSEA Fit Index,'' unpublished manuscript, University of British Columbia, Vancouver, BC, Canada.

Thurstone, L. L. (1947), *Multiple Factor Analysis,* Chicago: University of Chicago Press.

Verhage, Bronislaw J., Ugur Yavas, and Robert T. Green (1990), ''Perceived Risk: A Cross-Cultural Phenomenon?'' *International Journal of Research in Marketing,* 7 (4), 297–303.

Williams, Larry J. and Patricia J. Holahan (1994), ''Parsimony-Based Fit Indices for Multiple-Indicator Models: Do They Work?'' *Structural Equation Modeling,* 1 (2), 161–189.