2006

Examining the Cross-National Applicability of Multi-Item, Multidimensional Measures using Generalizability Theory

Srinivas Durvasula
Marquette University, srinivas.durvasula@marquette.edu

Richard G. Netemeyer
University of Virginia - Main Campus

J. Craig Andrews
Marquette University, craig.andrews@marquette.edu

Steven Lysonski
Marquette University, steven.lysonski@marquette.edu

Journal of International Business Studies, Vol. 37, No. 4 (2006): 469-483. DOI.

Examining the Cross-National Applicability of Multi-Item, Multi-Dimensional Measures Using Generalizability Theory

S Durvasula
College of Business Administration, Marquette University, Milwaukee, WI
R G Netemeyer
McIntire School of Commerce, University of Virginia, Virginia
J C Andrews
Department of Marketing, Marquette University, Milwaukee, WI
S Lysonski
Department of Marketing, Marquette University, Milwaukee, WI

Abstract

Establishing the applicability of multi-item measures is important for making valid inferences when testing theories cross-nationally. Typically, researchers have relied upon the tenets of classical measurement theory

(CT) using confirmatory factor model invariance testing to conclude that a unidimensional measure is applicable across countries. However, two important issues remain unresolved via CT techniques: (1) if the measure is found not to be invariant, CT tells us little as to why the measure varies across countries; and (2) if the measure is multi-dimensional, what factors affect its cross-national applicability? Our research seeks to address these issues and the cross-national measurement applicability of multi-dimensional scales via generalizability theory (GT). In this paper, we use a cross-national data set and simulated data sets to demonstrate the usefulness of GT to cross-national multi-dimensional measurement.

Introduction

Establishing the applicability of constructs and measures developed in the US to other countries continues to be a focus of cross-national marketing research and international business studies in general (Mullen, 1995; Balabanis et al., 2001; Steenkamp et al., 2001; Fu et al., 2004). All too often researchers have assumed that concepts and measures developed in one country are relevant in other countries without examining their cross-national applicability. The term 'applicability' refers to construct equivalence in which the operational definition and conceptual meaning of the construct are the same across countries. Applicability basically implies that the construct is expressed in a similar way in all countries of interest, and therefore has similar levels of reliability and validity. When the assumption of applicability is not verified, the probability of invalid cross-national inferences increases. That is, if the psychometric properties of a measure vary widely across countries, conclusions based on the scale may actually represent artifacts due to scale unreliability and lack of validity (Van de Vijver and Leung, 1997; Steenkamp and Baumgartner, 1998).

In addressing cross-national applicability of multi-item *unidimensional* measures, researchers have proposed methods based on classical measurement theory (CT) that use confirmatory factor analysis invariance testing (Steenkamp and Baumgartner, 1998). The primary finding from these methods is that a measurement scale is either country-specific or invariant cross-nationally. Although useful, CT methods leave two cross-national measurement issues unresolved. First, if a unidimensional scale is found not to be invariant, little is revealed as to what causes the measure to vary cross-nationally. Is the variation due to subjects' item responses? To country differences? Also, does lack of invariance negate the applicability of the scale cross-nationally? Second, CT will tell us little as to the causes of invariance for *multi-dimensional* scales. Could variation be due to a country by dimension interaction? To a subject within country by dimension interaction? Also, how do between-dimension correlations affect cross-national applicability?

A potential method to address these questions is generalizability theory (GT). GT recognizes that several sources (e.g., items, persons, countries, dimensions) can contribute to measurement error, and makes it possible to assess the combined and interaction effects of these sources across countries. Although applications of GT are beginning to appear in the marketing literature (Finn and Kayande, 1997), we are aware of only one study that uses GT to assess cross-national applicability (Sharma and Weathers, 2003), and this study focused only on a unidimensional scale, that is, the CETSCALE (Shimp and Sharma, 1987). For a unidimensional scale there are only a few sources of response variance (e.g., variance due to country differences, person differences, item differences, and the interaction between country and item). In contrast, multi-dimensional scales bring in three additional sources of variance:

1) dimension differences;
2) interactions between country and dimension; and/or
3) interactions between person and dimension.

Unlike unidimensional scales, applicability of multi-dimensional scales also requires assessment of discriminant validity among dimensions.

The purpose of our paper is to extend the work of Sharma and Weathers (2003) by demonstrating the usefulness of GT in cross-national measurement for multi-dimensional scales. We first provide an overview of GT and CT and the importance of establishing dimensionality. We then offer the procedures for conducting GT and the research expectations that require support for a multi-dimensional scale to exhibit cross-national applicability via GT. We next compare CT and GT using responses to a multi-dimensional advertising attitude (AA) measure collected across five countries. We also offer a GT simulation study, based on the AA data, which further shows GT's usefulness for assessing the cross-national applicability to multi-dimensional scales. We close our paper with a brief discussion with implications for GT and cross-national measurement.

Generalizability theory, classical theory, and the ANOVA model

Generalizability theory

The basic premise behind generalizability theory (GT) is that 'an investigator asks about the precision or reliability of a measure because he/she wishes to generalize from the observations in hand to some class of observations to which it belongs' (Cronbach et al., 1963: 144). For example, although a researcher may be interested in subjects' responses to a particular set of items on a particular occasion, he or she may be more interested in generalizing the observations over a set of measurement conditions. These conditions could be a 'universe' of similar items, a 'universe' of similar occasions, or a 'universe' of similar countries. In the domain of GT, a measurement condition is also called a *facet*. Facets are similar to factors in analysis of variance (ANOVA), in which conditions of each facet are analogous to levels of factors (Rentz, 1987; Shavelson and Webb, 1991; Finn and Kayande, 1997; Sharma and Weathers, 2003). The set of items of a scale is an example of a facet, and individual items of that scale would constitute conditions of that facet. Similarly, country and scale dimensionality are other facets where individual countries and dimensions represent conditions of these facets, respectively.

Facets can be categorized into two types: facets of generalization and facets of differentiation. *Facets of generalization* contribute to unwanted variation (i.e., measurement error). Consequently, the measurement instrument must be designed to minimize variance stemming from these facets. An example of a facet of generalization is a unidimensional scale such as the CETSCALE (Sharma and Weathers, 2003). To be generalizable, such a scale should minimize variance arising from differences among scale items. *Facets of differentiation* represent a set of objects which are to be compared for the study. For example, in any cross-national study it is typical to compare how subjects in various countries respond to a scale. Here, subjects and countries may serve as facets of differentiation. As not all subjects and countries are alike, responses from subjects within and across countries are likely to differ. So, differentiation facets such as subjects and countries contribute to variance that is desired or expected. In turn, measurement instruments should be designed to maximize variance from facets of differentiation.

Classical theory

As compared with GT, classical theory (CT) is less flexible. The main reason is that in CT persons are commonly assumed to be the facet of differentiation while a multi-item measurement scale is considered to be the facet of generalization. Therefore, applicability of CT becomes limited when the object of study is not persons, but rather stores, countries, ads, products, etc. (Rentz, 1987; Shavelson and Webb, 1991; Finn and Kayande, 1997). In contrast, GT is capable of estimating the applicability of measures when the object of study is not persons. Further, it is understood that what objects constitute the facets of differentiation and what factors constitute the facets of generalization can differ from study to study, depending on the research focus.

CT partitions observed score variance (σ_O^2) into two parts: that which is thought to be systematic, called true score variance (σ_t^2); and that which is thought to be random, called error variance (σ_e^2). Thus $\sigma_O^2 = \sigma_t^2 + \sigma_e^2$ (Shavelson and Webb, 1991; Nunnally and Bernstein, 1994). In most practical situations, error variance arises from multiple sources. Hence error estimates and estimates of reliability (i.e., ratio of true score variance to observed score variance) vary according to the data collection design. For example, in CT, *test–retest reliability* counts time variation as error but not variation due to item sampling, and *internal-consistency reliability* counts variation due to item sampling as error but not time variation. Thus CT treats each source of error independently and defines multiple reliability estimates by employing alternative definitions of what is true score and what is error. CT then precludes the estimation of combined effects of different sources of error or their potential interactions.

In contrast, GT recognizes that there may be many definitions of true and error scores. Moreover, multiple sources of error define the universe of generalization. Instead of asking how accurately observed scores reflect corresponding true scores, GT asks how accurately observed scores can be generalized to people's behavior in a defined universe of situations. The question of reliability evolves, then, into a question of generalization.

ANOVA as an analogy for GT and CT

GT is best described within the ANOVA framework, as GT is to measurement what ANOVA is to experimental research. We use ANOVA to identify and estimate the effects of important independent variables. Likewise, we use GT to identify and estimate the effects of important sources of error in a scale. In substantive research, the researcher manipulates independent variables while controlling/ignoring others. Likewise, GT methodically manipulates various facets of error, controls some, and ignores others.

As described by Shavelson et al. (1989), the comparison between GT and CT is similar to contrasting *factorial ANOVA* with *simple ANOVA*. When using simple ANOVA – akin to CT – variance is partitioned into 'between' and 'within' group variances. The 'between' group variance is viewed as systematic variance that contrasts groups from each other. The 'within' group variance is viewed as random and treated as an error because it diminishes the researcher's view of group differences. Likewise, CT partitions variance into true score and error variance. The true score variance is treated as systematic variance, associated with differences between the objects of measurement (i.e., persons). On the other hand, the error variance is viewed as random variance, which is unrelated to true score variance.

As compared with simple ANOVA, factorial ANOVA – akin to GT – assumes that multiple factors contribute to variance in the data. So, the total variance is partitioned into segments corresponding to each factor, the interactions among them, and to 'random error'. "Whereas CT partitions variance into only two sources, GT partitions variance into many sources, corresponding to systematic variance among the objects of measurement, to multiple error sources, and to their interactions" (Shavelson et al., 1989, p. 323).

The importance of dimensionality

Up to this point, we have discussed differences between GT and CT and some of the advantages that GT has over CT in establishing the cross-national applicability of measures. In our view, a primary advantage of GT is how it may be used to examine issues pertinent to the validity of dimensions within a multi-dimensional measure.

The importance of establishing dimensionality should not be understated (Clark and Watson, 1995; Netemeyer et al., 2003). To operationalize latent constructs, researchers often use composite scores by summing or averaging across items designed to measure the construct of interest. The uses of such scores are meaningful only if the items have acceptable dimensionality. Multi-dimensional scales, when treated as unidimensional (i.e., summed or averaged item composites), may result in interpretational ambiguities of the relationships among constructs in a test of theory. That is, if a construct is multi-dimensional, but all item scores are summed/averaged across dimensions into a single composite score and correlated with a criterion variable, such a correlation is at best ambiguous and, at worst, misleading.

Neuberg et al. (1997) offer an eloquent exposition of the potential problem of treating a multi-dimensional scale as if it were unidimensional by drawing an analogy with experimental manipulations and ANOVA. They suggest that a primary goal of experimentation is to create an unconfounded manipulation of an independent variable to accurately assess its effect on the dependent variable. If two constructs are actually being manipulated by a single experimental manipulation designed to manipulate one construct, that one construct's effect on the dependent variable cannot be accurately gauged, as disentangling its effect from the unwanted variation of the second construct is problematic. Similarly, survey researchers have the goal of developing a scale (or scale dimension) such that one construct is being assessed. The rationale behind unidimensionality is that the '…interpretation of any measure – whether it represents a trait, a mood, an ability, or a need – is clearest if only one dimension underlies the measure' (Neuberg et al., 1997: 1022). When only one dimension underlies a measure, that measure's correlation with a criterion is clearer. When more than one dimension exists, possibly suggesting that more than one trait/individual difference variable is being assessed, that measure's correlation with a criterion may be confounded. As such, establishing dimensionality is a necessary condition for internal consistency, construct validity, and theory testing. As we shall demonstrate shortly, GT can examine important dimensionality issues for cross-national measurement.

Application of GT to multi-dimensional scales

Designing a cross-national 'G-study'

Designing a *G*-study is similar to designing experiments. The researcher obtains measurements on the facet(s) of differentiation under various conditions of all relevant facets of generalization. Facets may be crossed or nested, random or fixed. In a single-country study, in which a sample of subjects evaluates a multi-item measurement scale, the subjects and items can be viewed as random facets, both of which are part of a crossed design. On the other hand, if the study is cross-national in nature, then subjects sampled in one country are not the same as those sampled in the other countries. In that case, subjects are nested within the country facet, which, in turn, is crossed with the items facet. A *coefficient of generalizability* may be computed for each facet or for various combinations of facets. This coefficient indicates the extent to which one can generalize the measurements to the universe of generalization.

Multi-facet analysis and coefficient calculations

Procedures for analyzing data in *G*-studies and computing *G*-coefficients are based on ANOVA. These procedures are also known as multi-facet analyses, and 'Proc Varcomp' in the SAS statistical package is often used. As in ANOVA, total score variance is partitioned into components of variance attributable to individual facets and their interactions. The variance components indicate which facets or interactions are contributing to substantial amounts of error and should be modified in subsequent studies. The variance components are also used to calculate generalizability coefficients. The formula for computing the *G*-coefficient is as follows (Shavelson and Webb, 1991):[Footnote1]

$$G = \frac{\sigma^2_{true\ score\ variance}}{\sigma^2_{true\ score\ variance} + \sigma^2_{error\ score\ variance}}$$

(1)

Partitioning true and error score variance for multi-dimensional scales

The components that contribute to error variance vary depending on whether *absolute error variance* $\left(\sigma^2_{Abs}\right)$ or *relative error variance* $\left(\sigma^2_{Rel}\right)$ is estimated. The absolute error variance is appropriate when the decision will be an absolute one *vs* established norms. For example, a country's score on an advertising scale might be considered valid if the score exceeds some predetermined norm. The decision is an absolute one in the sense that it does not depend on a ranking of other countries. On the other hand, the country's score might be considered valid if it exceeded another country's score. In this situation, the decision depends on a ranking so the relative error variance is more appropriate. The absolute error variance is at least as large as the relative error, and is usually larger because the relative error variance does not include variability attributable to the overall mean, whereas the absolute error variance does. CT uses relative error variance only because, by assumption, the mean score on parallel tests is always equal.

For a multi-dimensional scale, a key objective is to determine whether mean country scores can be generalized across items and dimensions. This results in a four-facet design involving persons (P), countries (C), dimensions (D), and items (I), in which the person facet is nested within the country facet. If the items of various dimensions are different, then the item facet is nested within the dimension facet. Whereas country and persons are the differentiation facets, items and dimensions are the generalization facets. Based on the design, it is possible to obtain variance (σ^2) components for C, $P{:}C$ (i.e., person nested within country), D, $I{:}D$ (i.e., item nested within dimension), $C \times D$ interaction, $I{:}D \times C$ interaction, $P{:}C \times D$ interaction, and E (i.e., error and all other confounded interactions). These variance components are used to compute the true score variance and error variance, where the true score variance is the sum of the variance components for the differentiation facets (i.e., P and C), where $\sigma^2_{true\,score\,variance} = \sigma^2_C + \sigma^2_{P:C}$. Equations (2) and (3) show how to compute the absolute and relative error score variances for multi-dimensional scales:

$$\sigma^2_{Abs} = \frac{\sigma^2_D}{n_D} + \frac{\sigma^2_{C\times D}}{n_D} + \frac{\sigma^2_{I:D}}{n_I \cdot n_D} + \frac{\sigma^2_{I:D\times C}}{n_I \cdot n_D} + \frac{\sigma^2_{P:C\times D}}{n_D} + \frac{\sigma^2_E}{n_I \cdot n_D}$$

(2)

$$\sigma^2_{\mathrm{Rel}} \frac{\sigma^2_{C\times D}}{n_D} + \frac{\sigma^2_{I:D\times C}}{n_I \cdot n_D} + \frac{\sigma^2_{P:C\times D}}{n_D} + \frac{\sigma^2_E}{n_I \cdot n_D}$$

(3)

where n_I is the number of items, and n_D is the number of dimensions.

Computing the reliability coefficient

The internal consistency reliability of a measurement scale can be obtained by treating person as the facet of differentiation and item as the facet of generalization. The reliability coefficient is calculated with only two sources of variation, person (P) and residual (E), representing error and the $P \times I$ interaction confound.

As P is the differentiation facet, the variance accounted for by this facet (σ_P^2) is the true score variance. As CT assumes parallel measurements, the item effect is assumed to be constant for all individuals: therefore, it is not included as a component when estimating the error variance. That is, the error variance is computed only from the variance accounted by the residual (σ_E^2). The formula for computing the reliability coefficient (R) is as follows, where σ_P^2 is the true score variance and σ_E^2/n_I is the error variance:

$$R = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_E^2/n_I}$$

(4)

As several facets of differentiation and generalization can be simultaneously considered for computing the *G*-coefficient, it is possible to compute either an overall *G*-coefficient for a multi-item multi-dimensional scale or a separate *G*-coefficient for each dimension. On the other hand, the *R* coefficient is based on the theory of individual differences only. Therefore there are separate estimates of *R*-values for each dimension of a multi-item scale and for each country.

Study 1

Research expectations and empirical demonstration

As discussed before, for a multi-dimensional scale where each dimension is measured by a separate (but correlated) scale, the persons facet is nested within the country facet and the items facet is nested within the dimension facet. Multi-facet analysis of the four-facet design produces variance components for country (C), dimension (D), person within country ($P{:}C$), item within dimension ($I{:}D$), country by dimension interaction ($C \times D$), country by item interaction ($I{:}D \times C$), dimension by person interaction ($P{:}C \times D$), and error. To support cross-national applicability, the following research expectations require support:

A: If the variance accounted for by between country differences (C, $C \times D$, $I{:}D \times C$) is smaller than variance accounted for by within-country differences ($P{:}C$, $P{:}C \times D$), then cross-national differences are less critical than within-country differences, enhancing generalizability.

B: If the variance accounted for by $I{:}D \times C$ is small, then the scale items are not country specific and generalizability is enhanced.

C: The variance accounted for by the dimension facet (D) should be smaller than the variance accounted for by the person and country facets. At the same time, as measures that tap different dimensions of a construct differ (as they should), the variance accounted for by the interaction of dimensions and persons within country (i.e., $P{:}C \times D$) should be significant, supporting the cross-national discriminant validity among dimensions and enhancing generalizability.

D: If the item facet ($I{:}D$) is relatively small, then the scale items are internally consistent, enhancing generalizability.

E: Assuming that country and subjects are the differentiation facets and items are the generalization facet, the overall generalizability coefficient (G) should be high. Such a finding implies that variation due to differences in countries and subjects can be generalized across the scale dimensions.

F: If a scale is truly multi-dimensional, the *G*-coefficient for such a scale would be significantly smaller than the *G*-coefficient computed for each dimension separately. The smaller *G*-coefficient for multidimensional scales is due to the dimension effect. This also supports discriminant validity of scale dimensions.

Methods

For illustration purposes, we applied GT to the advertising attitude (AA) construct. AA is a three-dimensional construct that has academic and practitioner importance for international marketing, as non-US advertising expenditures have increased by 60% since 1990 (Belch and Belch, 2004). It is also well known, that within countries, not all advertisements are equally attended to or liked by consumers (Onkvisit and Shaw, 1999). Such individual and advertising differences within countries can add variance that may reduce the ability of statistical models to find significant differences between countries. A theory that accounts for such measurement differences and their interactions is needed to provide a more accurate assessment of relative country differences. GT is such a theory.

The AA dataset measured cross-national attitudes toward advertising in five countries: New Zealand, Denmark, Greece, the United States, and India (sample sizes ranged from 87 to 179). In each country, non-probability samples of undergraduate students majoring in business (evenly divided by gender) provided responses to three dimensions of AA:

1. attitudes toward the institution of advertising (attitude–institution);
2. attitude toward the instrument of advertising (attitude–instrument); and
3. overall attitude toward advertising in general (attitude–general).

Whereas three-item scales measured attitude–institution and attitude–general, attitude–instrument was measured with four items (Durvasula et al., 1993). In each country, the survey was administered during class time.

An important prerequisite for any cross-national analysis is to first establish the *conceptual equivalence* of one's research materials. This is similar to the notion of ethnoconsumerism, which is the study of consumption from the point of view of the social groups or cultural groups being studied (Venkatesh, 1995). One method for establishing conceptual equivalence is to demonstrate translation equivalence, with a careful forward and backward translation of measurement items (Berry, 1980; Van de Vijver and Leung, 1997). Given that students in all countries but Greece spoke and read English fluently, only the Greek version of the survey required forward and backward translation. Employing the standard questionnaire translation procedure, the Greek survey was translated into Greek with the aid of two bilingual experts.

Although student samples have been criticized for a potential lack of sample representativeness, they are appropriate for the present research for two reasons. First, our research can be classified as a comparative or theoretical test of the validity of attitudes in two or more countries. This type of research favors samples that ensure that any observed differences in the constructs are not due to sample differences. Thus homogeneous samples that control for demographic characteristics are desired, and non-probability samples are acceptable (Whitman et al., 1999; Reynolds et al., 2003).

Given that multi-facet analysis requires a balanced design, random sampling within country was used to achieve an equivalent sample size of 87 in each country for analysis purposes. Thus, following pooling of data across subjects (87 per country), countries (five), and 10 items (three items for two dimensions of AA and four items for one dimension of AA), there were a total of 4350 (87 × 5 × 10) observations in the analysis. The multi-facet analysis is a nested one, where subjects are nested within country and items are nested within dimension. Hence there are no subject and item facets directly. The person:country $(P{:}C)$ facet reflects the effect of persons within each country, and the item:dimension $(I{:}D)$ facet reflects the effect of items within each dimension. As a result of the nested design, several interaction effects could not be measured directly.[Footnote2]

Results: CT and GT comparisons

CT results

To demonstrate potential differences between GT and CT, we first applied the CT technique of multi-group hierarchical model invariance testing via confirmatory factor analysis (CFA) to the data. We used the framework suggested by Steenkamp and Baumgartner (1998) with sample sizes of $n = 87$ per country – the same samples used for the GT analyses that follows.

The first model estimated in the hierarchy is the configural invariance model ($\chi^2 = 329.92, \mathrm{df} = 160$). Configural invariance suggests that a *similar* pattern of item-to-dimension loadings and correlations among dimensions exists across countries. Thus configural invariance is supported if all items to their respective dimensions have significant loadings, the correlations among the three AA dimensions are significantly different from 1 (Anderson and Gerbing, 1988), and model fit indices are at adequate levels. Results revealed significant factor loadings for all items across all countries, and correlations among AA dimensions were significantly different from 1. Three fit indices – the non-normed fit index (NNFI), the comparative fit index (CFI), and the root mean squared error of approximation (RMSEA) – were used to assess model fit. Levels of 0.95 and above have been advocated for NNFI and CFI, and levels of 0.06 and below have been advocated for RMSEA (Hu and Bentler, 1999). These fit indices were at marginal levels (NNFI=0.89; CFI=0.92; RMSEA=0.11). In sum, the configural invariance model was supported on two criteria, but only marginally supported on the third.

The next model estimated is the metric invariance model ($\chi^2 = 372.34, \mathrm{df} = 188$). To test for metric invariance, item factor loadings are constrained to be the same across countries. Although the indices for this model (NNFI=0.90; CFI=0.92; RMSEA=0.11) are similar to those for the configural invariance model, of importance for testing metric invariance is the difference in χ^2 fit between the configural and metric invariance. The difference was significant (χ^2 diff $= 42.42, \mathrm{df} = 28,\ P < 0.05$), suggesting that not all item loadings to their respective dimensions were statistically invariant across the five countries. Still, given that a χ^2 difference of 41.34 is required for a difference at the 0.05 level, and the difference observed was just minimally above that (i.e., 42.42), some evidence for the metric invariance of the AA items is found.

The third model specifies invariant factor variances and covariances of the AA dimensions across countries, as well as invariant item factor loadings ($\chi^2 = 434.06, \mathrm{df} = 212$). This model did differ statistically from the configural invariance model (χ^2 diff $= 104.14, \mathrm{df} = 52,\ P < 0.01$), while witnessing only slight depreciations in NNFI (0.89), CFI (0.90), and RMSEA (0.12). This suggests that the relationships (correlations) among AA dimensions likely differ across samples.

Finally, we estimated a model specifying invariant item loading error variances, invariant factor variances and covariances, and invariant item factor loadings ($\chi^2 = 882.87, \mathrm{df} = 252$). This model significantly differed in fit from the configural invariance model (χ^2 diff $= 552.95, \mathrm{df} = 92,\ P < 0.01$), and did have substantially lower fit indices (NNFI=0.75, CFI=0.72, and RMSEA=0.20). However, since a focus of our paper is to compare CT and GT, finding support for error variance equivalence is not essential (Steenkamp and Baumgartner, 1998). In sum, CT results suggest that the AA measure has some level of cross-national applicability across the five countries.[Footnote3]

GT results

The top portion of Table 1 shows results of the GT multi-facet analysis at the cross-national level for the advertising attitude (AA) data. For each component, the estimated variance as well as the percentage of total variance accounted for by that component is given. It is clear from the table that the persons within country facet accounted for the largest variance (26.41%). This effect is a result of individual differences among subjects within each country. The size of this effect is over seven times the effect due to cross-national (or between-group) differences of 3.87%. Further, by adding the interactions of persons within country facet with other facets, we obtain the proportion of variance attributed to *within-country effects* of 37.68% (i.e., 26.41%+11.27%). On the other hand, the proportion of variance accounted for by all *between-country (or between-group) effects* is only 6.68% (i.e., 3.87%+2.46%+0.35%). So, the variance introduced by individual differences is about six times larger than that introduced by between country differences. Responses within countries are so diverse that cross-national differences add only slightly to the variance already explained by personal (or within-country) differences. Therefore, for advertising attitudes in our study context, cross-national differences are less significant than within-country differences, supporting research expectation **A** and cross-national generalizability.[Footnote4]

Table 1 Multi-facet analysis of cross-national advertising data

Source of variance	*Variance component*	*% of total variance*
Country	0.11	3.87
Dimension	0.60	21.13
Person: country	0.75	26.41
Item: dimension	0.00	0.00
Person: country × dimension	0.32	11.27
Item: dimension × country	0.01	0.35
Country × dimension	0.07	2.46
Other interactions and error	0.98	34.51
Total	2.84	100.00
Attitude–institution dimension		
Country	0.16	7.08
Person: country	1.24	53.78
Item	0.00	0.00
Country × item	0.00	0.15
Other interactions and error	0.90	39.03
Total	2.31	100.00
Attitude–instrument dimension		
Country	0.19	8.97
Person: country	0.74	33.82
Item	0.10	4.37
Country × item	0.02	0.84
Other interactions and error	1.13	52.00
Total	2.17	100.00
Attitude-toward-advertising-in-general dimension		
Country	0.16	6.47
Person: country	1.57	65.29
Item	0.02	0.68
Country × item	0.02	0.89
Other interactions and error	0.64	26.67
Total	2.40	100.00

From the top portion of Table 1, it also is clear that *dimensionality issues* contributed significantly to the overall variance. Among them, the variance accounted for by the dimension facet is 21.13%. This is due to overall differences in responses to the three advertising measures across the countries and subjects. Although the proportion of variance accounted for by the dimension facet is rather high, it is still smaller than the variance accounted for by within-country differences. Also, one would expect that, *a priori,* this proportion would be nontrivial because the three dimensions tap different aspects of advertising attitudes, supporting research expectation **C**.

In contrast, *measurement issues* (i.e., $I:D$ and $I:D \times C$) contribute very little to overall variance. Among them, the *items within dimension* facet $(I:D)$ is small, representing approximately 0% of the total variance. This suggests that the items within each dimension are quite homogeneous, exhibiting only minor differences. This result is analogous to finding high-scale reliabilities in CT analyses. Further, the relatively small *items within dimensions by countries* interaction $(I:D \times C)$ suggests that scale items are not country specific, offering support for research expectations **B** and **D**. Note that finding a small $I:D \times C$ is similar to finding support for metric invariance in CT analyses. However, where CT analyses found only marginal support for metric invariance based on the χ^2 difference tests, GT analyses provide stronger support.

To further understand the relative impact of various facets on total variance, we analyzed the data for each advertising attitude separately. If analyzing multi-dimensional data is similar to performing, then analyzing data for each dimension separately is similar to performing multiple ANOVAs or performing CT analyses on unidimensional scales. The effective sample size then is 1305 (i.e., 87 subjects × five countries × three items) for attitude – institution and attitude – general. For attitude – instrument, the effective sample size is 1740 (i.e., 87 subjects × five countries × four items). As indicated in the bottom portion of Table 1, across the three AA dimensions, the proportion of variance accounted for by the persons within country facet is much larger than that found for the country facet. On the other hand, the variance accounted for by measurement issues (i.e., *item* effect and *countries by items within dimensions* interaction) is consistently small across the dimensions, reflecting the small item level differences. In sum, these results are consistent with the results of the pooled data that are presented in the top portion of Table 1. The between-country differences in this study are much smaller than individual differences within each country, supporting research expectation **A**, and enhancing cross-national generalizability.

Results of multi-facet analysis also can be used to compute the *G*-and reliability coefficients. The *G*-coefficient indicates the extent to which advertising measures can be generalized; the reliability coefficient indicates the extent to which the advertising measures are internally consistent. Also, whereas the *G*-coefficient is a summary coefficient at the cross-national level, the reliability coefficient is computed for each country and advertising dimension separately. Table 2 shows the results. For the *G*-and reliability coefficients, the magnitude of the coefficient can be interpreted in the same manner. Across the AA dimensions, the *G*-coefficients for countries and subjects-within-countries are moderate, but above 0.7 – a level indicative of adequate generalizability (Rentz, 1987).

Table 2 Generalizability coefficients for the advertising data

Dimension	$G_{Abs}\ (SC)$	$G_{Rel}\ (SC)$	Reliability
Attitude–institution			
Cross-national sample	0.73	0.74	
New Zealand			0.81
Denmark			0.89
Greece			0.63
United States			0.88
India			0.73
Attitude–instrument			
Cross-national sample	0.75	0.76	
New Zealand			0.70
Denmark			0.72
Greece			0.80

United States			0.76
India			0.66
Attitude-toward-advertising-in-general			
Cross-national sample	0.88	0.89	
New Zealand			0.89
Denmark			0.91
Greece			0.88
United States			0.93
India			0.81

We computed *G*-coefficients for the multi-dimensional scale as well as for each of the three AA dimensions using equations (1), (2) and (3). The *G*-coefficient for the multi-dimensional scale based on absolute error (G_{Abs}) is 0.65, and is 0.77 based on the relative error (G_{Rel}). In contrast, the *G*-coefficient (G_{Abs}), when computed for each AA dimension separately, ranges from 0.73 to 0.88 (see Table 2). The *G*-value for the multi-dimensional scale is lower because of the significant dimension effect. This suggests that the three scales tap different dimensions of AA. A large dimension effect, when coupled with a small item effect (as we found), indicates that the scale dimensions possess discriminant validity, supporting research expectation **F**.

The overall G_{Rel} for the multi-dimensional scale is 0.77. Although this value is not as sensitive to the dimension effect, the size of this coefficient (>0.7) provides support for the cross-national applicability of the AA scale. (See research implication **E**.)

From Table 2, reliability coefficients provided at the country level also are supportive of the three advertising measures, as most of the reliability coefficients (13 of 15) are above 0.70 (Nunnally and Bernstein, 1994). More importantly, the small number of items in each dimension coupled with the high average inter-item correlations within dimension (0.57 for attitude – institution, 0.41 for attitude – instrument, and 0.72 for attitude toward advertising in general across countries) supports the internal consistency of each dimension (Netemeyer et al., 2003). In sum, results of multi-facet analyses provide support for the cross-national applicability of AA measures.

Study 1 summary

Overall, Study 1 results show that the advertising attitude measure (AA) consisting of three dimensions – attitude institution, attitude instrument, and attitude general – shows an adequate level of cross-national applicability via CT, and a stronger level via GT.[Footnote5]

Although these results demonstrate the usefulness of GT in establishing the cross-national applicability of multi-dimensional scales, they tell us little about the sensitivity of GT coefficients, and hence the dimensionality of measures when correlations among dimensions differ. It may be useful to the cross-national researcher to find out how such varying correlations affect the different facets (e.g., C, $P{:}C$, $I{:}D$) and their interactions (i.e., $P{:}C \times D$, $I{:}D \times C$, and $C \times D$). Study 2 addresses this issue.

Study 2

Intuitively, we identify a condition that may affect the sensitivity of GT facets and their interactions, namely, correlations among AA dimensions. We examine correlations among dimensions for two reasons. First, and as previously noted, dimensionality is an important aspect of establishing measure validity, and hence cross-national applicability. As correlations among dimensions within a multi-dimensional scale increase toward unity, discriminant validity (and dimensionality) becomes threatened as two dimensions may, in fact, be tapping the same construct. Second, our CT analysis showed likely differences among the correlations of the dimensions of the AA measures. Thus, based on our AA data, we conducted simulations to examine the effect of increasingly high correlations among dimensions (approaching unity) on various facets and their interactions.

Simulation procedures

Correlations among AA dimensions of attitude–institution (AI) and attitude–general (AG), as well as attitude–institution (AI) and attitude–instrument (AR), were gradually increased to unity across all countries. The rationale for increasing AA dimension correlations to unity is threefold. First, across the five countries, the average correlations among AA dimensions (AI–AG, AI–AR, and AR–AG) are fairly high, ranging from 0.54 to 0.73 with an overall average of 0.67. As such, the simplest way to determine the sensitivity of various facets and their interactions is by analyzing the data when AA dimensions correlate well above these levels, that is, unity. Second, when the correlation between any two AA dimensions is unity, these two dimensions lack discriminant validity. It is interesting to determine which facets and their interactions are sensitive to a lack of discriminant validity. Third, when manipulating correlations among AA dimensions, it is important to keep item means and item variances constant, such that any changes in facet coefficients and their interactions are not due to differences in item means or variances. The most effective way to accomplish this is by setting correlations among dimensions to unity.

Table 3 presents four simulations. In the first two simulations, corresponding to rows 2 and 3 of Table 3, we randomly selected two AA dimensions, AI and AG, and set their correlation to unity. To keep the number of simulations to a minimum, AI–AG correlation was set to unity initially in *two randomly selected countries*, Denmark and India. The variance component estimates corresponding to this simulation are in the second row of Table 3. Then, we set AI–AG correlation to unity in the other three countries (New Zealand, Greece, and the US) as well. Results pertaining to this simulation are in the third row of the table. In the next set of simulations, while keeping the AI–AG correlation set to 1, we set AI–AR correlation to 1, again in Denmark and India initially (see the fourth row). Finally, AI–AR correlation is set to one in all five countries. These results are in the fifth row of Table 3.[Footnote6]

Table 3 Results of simulations

Simulated data	Variance accounted for by								% Var due to between-group effects	% Var due to within-group effects
	C	$P: C$	D	$I: D$	$P: C \times D$	$I: D \times C$	$C \times D$	Error		
Original data	0.106	0.753	0.602	0.002	0.316	0.011	0.071	0.979	6.62	37.64
Corr (AI–AG)=1 in Denmark, India	0.106	0.807	0.601	0.002	0.241	0.011	0.072	1	6.65	36.9
Corr(AI–AG)=1 in all five countries	0.105	0.813	0.602	0.002	0.181	0.01	0.072	1.054	6.59	35.01
Corr(AI, AR, AG)=1 in Denmark,										
India Corr(AI, AR)=1 in other countries	0.101	1.13	0.596	0.004	−0.007	0.011	0.076	0.921	6.64	39.65
Corr(AI, AR, AG)=1 in all five countries	0.096	1.453	0.587	0.007	−0.239	0.013	0.081	0.844	6.69	42.72

Note: The variance components were estimated by the MIVQUE algorithm of PROC VARCOMP in SAS, which is the recommended algorithm for both balanced and unbalanced designs (Rao, 1971). Theoretically, the variance component values should be non-negative because they are assumed to represent the variance of a random variable. Nevertheless, when using the MIVQUE method, it is not unusual to find that some variance components estimates are negative. In our case, the variance component for $P: C \times D$ interaction was negative in row 5. Other algorithms, such as maximum likelihood, would not have produced negative estimates. We used MIVQUE so that our results are consistent with those reported in Tables 1 and 2. We have also analyzed the data using maximum likelihood method and found that the overall interpretation of results remained the same. Further, the SAS manual suggests that it is common practice to treat negative variance components as if they are zero.

Simulation results

As previously noted, if very high correlations were to exist among AA dimensions, then those dimensions would lack discriminant validity. Table 3 indicates that, when this is the case, the *persons within countries by dimension* interaction $(P\colon C \times D)$ becomes smaller. Thus the contribution of dimensionality issues $(D,\ P\colon C \times D,\ C \times D)$ to overall variance also becomes smaller. Even though the variance contributed by the *P*:*C* facet becomes larger, the relative contribution of within-group effects vis-à-vis between-group effects remains the same. The contribution of measurement issues, or choice of measures $(I\colon D,\ I\colon D \times C)$, to overall variance is largely unaffected. Thus, the contribution of the dimensionality issues in general, and $P\colon C \times D$ interaction in particular, serves as an indicator of the presence or absence of discriminant validity among AA dimensions. Further, in the absence of discriminant validity among dimensions, a small $P\colon C \times D$ interaction effect implies that, across the countries, the three AA dimensions had the same meaning to the sampled persons.

For comparison purposes, in CT analyses, lack of discriminant validity can be statistically detected by finding out whether confidence intervals of correlations among AA dimensions contain a value of 1. While CT analysis of our original AA data supports discriminant validity, results show that in one country the correlation between two AA dimensions was fairly close to one (0.83). GT analysis of the original data, on the other hand, shows that $P\colon C \times D$ interaction is very much different from zero, more clearly implying that the three AA dimensions do possess discriminant validity.[Footnote7]

Discussion

Summary and conclusions

Measurement instruments developed in the US are increasingly being applied in different countries, raising a critical question of interest. Do these instruments apply to other countries? Until now, the answers to these questions have been supplied primarily by classical theory (CT) techniques. Confirmatory factor analysis has been the most common way to test the applicability of scales when they are used in other countries. If a scale is not cross-nationally applicable, a key question is why? Is it the items, subjects, dimensions, countries, or persons that contribute to variation in responses? Generalizability theory (GT) is an appropriate alternative for answering such questions (Finn and Kayande, 1997). Proponents of GT argue that application of CT in this case is limited because it does not distinguish among alternative sources of response variance.

For illustration purposes, we used the multi-dimensional advertising attitudes scale (AA). Results indicate that the AA dimensions have acceptable internal consistency estimates at the country level, but across the countries the dimensions possess only moderate-sized generalizability coefficients. A further investigation of the sources of variance revealed that most of the variance is due to within-country differences among subjects. Response variances between countries are relatively much smaller. For consumer behavior scales in general, such results imply that future research efforts should focus more on within-country differences than on cross-national differences.

We also found (for AA) that the dimension effect is larger than the item effect or the country effect, though smaller than the within-country person effect. This suggests there is sizeable variance in responses across scale dimensions, and supports the notion that the dimensions are related, but still possess discriminant validity. This conclusion is also supported by the size of the *persons within countries by dimensions* interaction, which is significantly different from zero. A relatively small *items within dimensions* effect indicates that the scale items are relatively homogeneous. Further, a small *items within dimensions by countries* $(I\colon D \times C)$ interaction implies that the scale items appear to have the same meaning across the countries. Such results suggest that those scales can be used for making meaningful cross-national comparisons. In contrast, a larger $I{:}D \times C$ interaction would have implied that the scale lacks construct equivalence.

Implications and advantages of GT

Though our illustration shows that the GT methodology works well for multi-dimensional constructs, certain issues relative to CT need to be addressed. For instance, the GT methodology employs ANOVA framework in partitioning the response variance. Hence to achieve a balanced design the sample sizes across the countries must be the same. We accomplished this objective by identifying the smallest sample size in any country as the basis and taking random samples from other countries to match that size. Likewise, for multi-dimensional scales, it is preferred to have an equal number of items per scale dimension. This requirement is difficult to satisfy. For example, the AA scale had an unequal number of items across scale subdimensions. In that case, it is still possible to perform a multi-facet analysis and estimate the impact of alternative variance sources (e.g., country, subject, item, and dimension). Still, to compute the overall generalizability coefficient, a random sample of items needs to be taken from some of the scales in order to have an equal number of items across all scale dimensions.[Footnote8]

For those who wish to assess cross-national applicability of measures, both CT and GT approaches offer advantages and disadvantages. CT follows a well-developed invariance framework for scale evaluation that uses statistical tests. However, the χ^2 difference test used in invariance testing is often criticized for its sensitivity to sample size effects, which could result in rejecting a reasonably good cross-nationally applicable scale. Also, CT analyses assume that persons are the object of measurement. As such, CT is not as useful if the objects of measurement are countries, dimensions, or items.

In contrast, GT analysis is easier to understand because of its similarity to the ANOVA framework. For cross-national evaluation of multi-dimensional scales, the factors that contribute to overall response variance are countries, persons (within countries), dimensions, items (within dimensions), and their interactions. Standard statistical packages such as SAS and SPSS are available for performing variance decomposition analysis. Although not as statistically rigorous as CT-based invariance testing, GT offers more diagnostics for scale evaluation. For example, in the case where a CFA model produced poor or marginal fit for a measurement instrument (as was the case for our AA data), GT analyses can help to determine whether the major contributors of this lack of fit are country and higher-order country effects, persons and higher-order person effects, dimensions and higher-order dimension effects, or measurement issues related to choice of scale items. GT analyses are also advantageous when planning future cross-national studies. Sensitivity analysis based on the GT framework would indicate what sample size to use and how many scale items are needed to achieve a certain level of scale generalizability (Sharma and Weathers, 2003). Also, GT analyses are the most appropriate alternative for scale evaluation when data are collected on different occasions and by different interviewers. In that case, we would be able to find out the relative contributions not only of items and dimensions, but also of measurement occasions and interviewers. This is not possible in CT.

Although CT procedures for evaluating cross-national applicability of multi-dimensional scales are widely discussed in the literature, our study is the first one to examine this issue based on GT. When performing GT analysis, cross-national applicability of multi-dimensional scales requires that country-related effects and dimension effects should contribute less to the overall response variance than person-related effects. All told, then, GT provides more diagnostic information to assess the cross-national applicability of measurement scales than CT. By developing a procedure based on the GT for identifying alternative sources of response variance in cross-national research, our research offers an alternative and/or a complementary method to CT for establishing cross-national validity of measurement scales.

Ultimately, though, regardless of the technique used (whether CT or GT based) to assess cross-national scale applicability, the scale's development should be based on sound theoretical reasoning. Any justification for the scale's cross-national use should also be based on sound theoretical reasoning (i.e., the construct is substantively applicable and important cross-nationally), and any cross-national modification to the scale should be based on both strong theory and empirical support. Clearly, cross-national comparisons will not be valid if the scale possesses a different substantive meaning (i.e., conceptually not equivalent) across countries (Berry, 1980; Venkatesh, 1995) in the first place.

Notes

1. The procedures for performing multi-facet analysis are as follows. Suppose we have cross-national data on a multidimensional construct XYZ. We create a data file with separate columns for the variables person, country, item, dimension, and response to XYZ. Following are examples of how variance components such as those shown in Table 1 can be estimated in SPSS and SAS. Note that items are nested within dimension and persons are nested within country.Implementation in SPSS:VARCOMPXYZ BY country person item dimension/RANDOM=country person item dimension/METHOD=MINQUE (1)/DESIGN country dimension item (dimension) person (country) country × item (dimension) dimension × person(country) country × dimension.**Implementation in SAS:**PROC VARCOMP DATA=temp METHOD=MIVQUE0;CLASS country dimension person item;MODEL XYZ=country dimension person(country) item(dimension)dimension × person(country) country × item(dimension);RUN;
2. As there is only one observation per cell, the interactions that could not be estimated separately are $P{:}C \times C$, $I{:}D \times D$, $P{:}C \times I{:}D$, $P{:}C \times C \times D$, $I{:}D \times D \times C$, $P{:}C \times D \times I{:}D$, and $C \times P{:}C \times D \times I{:}D$. Thus, the error variance estimate includes variation due to these interactions.
3. As noted by Horn (1991: 125), metric invariance is a ‘reasonable ideal … a condition to be striven for, not one expected to be fully realized.’ Steenkamp and Baumgartner (1998) also note that, in practice, full metric invariance across numerous groups (countries) is unlikely. Marsh (1994) further notes that, as sample size increases, the probability of finding invariant models based on χ^2 difference tests decreases, and one should look at changes in other fit indices (CFI, NNFI, RMSEA) in assessing model invariance. When there is little depreciation in these indices, some evidence of ‘practical’ invariance exists. This is what we found in testing the AA scales.
4. As Steenkamp and Baumgartner (1998) suggest, cross-national mean differences should be examined only after establishing the cross-national scale applicability. For example, if a country's culture dimensions would dictate differences in cross-national construct of great interest, then after establishing that the scale to measure the construct is reliable and valid across countries, testing for mean level differences can be done. This is what Lenartowicz and Johnson (2003) illustrated in their paper. In GT, a relatively large dimension effect (*D*) or a country by dimension interaction $(C \times D)$ indicates potential construct mean differences across countries.
5. When comparing CT and GT it is important to note that CT is more rigorous in its strict adherence to χ^2 difference tests for establishing invariance. Still, the results generally converge at a ‘practical’ level when alternative fit indices such as GFI and CFI are also considered in CT for scale evaluation (Marsh, 1994).
6. Had we randomly selected New Zealand, Greece, and the US rather than Denmark and Greece, only the variance component estimates of the second and the fourth rows of Table 3 would differ, but these components would remain the same for the third row and the last row. Also, the overall interpretation of the results would not be affected by randomly selecting a different set of countries for the first simulation.

7. We also attempted to conduct CFA simulations (CT analyses) as well that could be comparable to our GT simulations. However, we could not obtain converged solutions when setting correlations among dimensions to unity. When correlations among dimensions are about unity, the covariance input matrix became 'not positive definite.' The rows of this input matrix are then linearly dependent on each other, thereby making parameter estimation impossible. Though we could not perform CFA analysis of simulated data, we would expect the results of such an analysis to show that the dimensions that have high correlations would fail the discriminant validity test.
8. As also shown by Sharma and Weathers (2003), differing sample sizes did not appreciably affect GT. We conducted our GT analyses with the original sample sizes across countries ($n = 87 – 179$) with essentially the same findings as those with our $n = 87$ across-countries samples. For example, the proportion of variance accounted for by within-country effects is between three and four times as large as the proportion of variance due to country effects. Dimensionality issues also contributed significantly to the overall variance. The variance accounted for by the dimension differences is 21.6%, which is close to what we reported for the balanced sample. Similar to what we found for the balanced sample, item differences accounted for a negligible percent of the total variance (0.4%).

References

Anderson, J.C. and Gerbing, D.W. (1988) 'Structural equation modeling in practice: a review and recommended two-step approach', Psychological Bulletin 103 (3): 411–423.

Balabanis, G., Diamantopoulos, A., Mueller, R.D. and Melewar, T.C. (2001) 'The impact of nationalism, patriotism, and internationalism on consumer ethnocentric tendencies', Journal of International Business Studies 32 (1): 157–175.

Belch, G.E. and Belch, M.A. (2004) Advertising and Promotion: An Integrated Marketing Communications Perspective, 6th edn, McGraw-Hill Irwin: Boston.

Berry, J.W. (1980) 'Introduction to methodology', in H.C. Triandis and J.W. Berry (eds.), Handbook of Cross-Cultural Psychology: Methodology, Vol. 2, Allyn & Bacon: Boston, pp: 1–28.

Clark, L.A. and Watson, D. (1995) 'Constructing validity: basic issues in objective scale development', Psychological Assessment 7 (3): 309–319.

Cronbach, L.J., Rajarathnam, N. and Gleser, G.C. (1963) 'Theory of generalizability: a liberalization of reliability theory', British Journal of Statistical Psychology 16 (November): 137–163.

Durvasula, S., Andrews, J.C., Lysonski, S. and Netemeyer, R.G. (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.

Finn, A. and Kayande, U. (1997) 'Reliability assessment and optimization of marketing measurement', Journal of Marketing Research 34 (May): 262–275.

Fu, P.P, Kennedy, J., Tata, J., Yukl, G., Bond, M.H., Peng, T., Srinivas, E.S., Howell, J.P., Prieto, L., Koopman, P., Boonstra, J.J., Pasa, S., Lacassagne, M., Higashide, H. and Cheosakul, A.A. (2004) 'The impact of societal cultural values and individual social beliefs on the perceived effectiveness of managerial influence strategies: a Meso approach', Journal of International Business Studies 35 (4): 277–283.

Horn, J.L. (1991) 'Comments on 'issues in factorial invariance'', in L.M. Collins and J.L. Horn (eds.), Best Methods for the Analysis of Change, American Psychological Association: Washington, DC, pp: 114–125.

Hu, L. and Bentler, P.M. (1999) 'Cutoff criteria for fit indices in covariance structure analysis: conventional criteria *versus* new alternatives', Structural Equation Modeling 6 (1): 1–55.

Lenartowicz, T. and Johnson, J.P. (2003) 'A cross-national assessment of the values of Latin American managers: contrasting hues or shades of gray', Journal of International Business Studies 34 (4): 266–281.

Marsh, H.W. (1994) 'Confirmatory factor analysis models of factorial invariance: a multi-faceted approach', Structural Equation Modeling 1 (1): 5–34.

Mullen, M.R. (1995) 'Diagnosing measurement equivalence in cross-national research', Journal of International Business Studies 26 (3): 573–596.

Netemeyer, R.G., Bearden, W.O. and Sharma, S. (2003) Scaling Procedures: Issues and Applications, Sage: Thousand Oaks, CA.
Neuberg, S.L., West, S.G., Thompson, M.M. and Judice, T.N. (1997) 'On Dimensionality, discriminant validity, and the role of psychometric analyses in personality theory and measurement: reply to Kruglanski *et al.*'s (1997) defense of the Need For Closure Scale', Journal of Personality and Social Psychology 73 (5): 1017–1029.
Nunnally, J.C. and Bernstein, I.H. (1994) Psychometric Theory, McGraw-Hill: New York.
Onkvisit, S. and Shaw, S.S. (1999) 'Standardized international advertising: some research issues and implications', Journal of Advertising Research 39 (6): 19–24.
Rao, C.R. (1971) 'Minimum variance quadratic unbiased estimation of variance components', Journal of Multivariate Analysis 1 (December): 445–456.
Rentz, J.O. (1987) 'Generalizability theory: a comprehensive method for assessing and improving the dependability of marketing measures', Journal of Marketing Research 24 (February): 19–28.
Reynolds, N.L., Simintiras, A.C. and Diamantopoulos, A. (2003) 'Theoretical justification of sampling choices in international marketing research: key issues and guidelines for researchers', Journal of International Business Studies 34 (1): 80–89.
Sharma, S. and Weathers, D. (2003) 'Assessing generalizability of scales used in cross-national research', International Journal of Research in Marketing 20 (3): 287–295.
Shavelson, R.J. and Webb, N.M. (1991) Generalizability Theory: A Primer, Sage Publications: Newbury Park, CA.
Shavelson, R.J., Webb, N.M. and Rowley, G.L. (1989) 'Generalizability theory', American Psychologist 44 (6): 922–932.
Shimp, T.A. and Sharma, S. (1987) 'Consumer ethnocentrism: construction and validation of the CETSCALE', Journal of Marketing Research 24 (August): 280–289.
Steenkamp, J.B.E.M., Batra, R. and Alden, D.L. (2001) 'How perceived brand globalness creates brand value', Journal of International Business Studies 34 (1): 53–65.
Steenkamp, J.B.E.M. and Baumgartner, H. (1998) 'Assessing measurement invariance in cross-national consumer research', Journal of Consumer Research 25 (June): 78–90.
Van de Vijver, F.J.R. and Leung, K. (1997) 'Methods and data analysis of comparative research', in J.W. Berry, Y.P. Poortinga and J. Pandey (eds.), Handbook of Cross-Cultural Psychology, Volume One: Theory and Method, Allyn & Bacon: Needham Heights, MA, pp: 247–300.
Venkatesh, A. (1995) 'Ethnoconsumerism: a new paradigm to study cultural and cross-cultural consumer behavior', in J. Costa and G. Bamossy (eds.), Marketing in a Multicultural World: Ethnicity, Nationalism, and Cultural Identity, Sage: Thousand Oaks, CA, pp: 26–67.
Whitman, M.E., Townsend, A.M. and Hendrickson, A.R. (1999) 'Cross-national differences in computer-use ethics: a nine-country study', Journal of International Business Studies 30 (4): 673–687.

Acknowledgements

The authors would like to thank the three anonymous reviewers and Professor Guliz Ger, the Departmental Editor of *JIBS*, for their helpful comments on earlier versions of this manuscript.