Relaxing Measurement Invariance in Cross-National Consumer Research Using a Hierarchical IRT Model

MARTIJN G. DE JONG
JAN-BENEDICT E. M. STEENKAMP
JEAN-PAUL FOX*

With the growing interest of consumer researchers to test measures and theories in an international context, the cross-national invariance of measurement instruments has become an important issue. At least two issues still need to be addressed. First, the ordinal nature of the rating scale is ignored. Second, when few or no items in the confirmatory factor analysis (CFA) exhibit metric and scalar invariance across all countries, comparison of results across countries is difficult. We solve these problems using a hierarchical IRT model. An empirical application is provided for susceptibility to normative influence, using a sample of 5,484 respondents from 11 countries on four continents.

Consumer researchers are becoming increasingly interested in testing their measures and theories in an international context (Bagozzi 1994; Durvasula et al. 1993; Wong, Rindfleisch, and Burroughs 2003). It is in this vein that Monroe (1993, v) urges consumer behavior researchers "to move beyond the relative security of our own backyards and investigate issues relative to consumption on an international basis." Consider the following substantive questions that consumer researchers may want to address:

- A consumer researcher is interested in testing whether materialism is largely a U.S. construct ("emic") or a pancultural construct ("etic"). To address this question, s/he wants to test the nomological relations between this construct and antecedents, consequences, and concurrent constructs as identified in U.S. research (Richins 1994; Richins and Dawson 1992) in other cultures.
- Cultural theory (Schwartz 1994) predicts that in countries with a high degree of embeddedness, the subjective norm is more important than a person's own attitude in shaping consumer behavior, while the converse is expected to be true in countries with a high degree of autonomy. Is this truly the case? Or are personal opinions the key driver of behavior, across cultures? What are the implications for decision theory and purchase models?
- Ever since Mick's (1996) seminal article, consumer researchers are well aware of the biasing effects of socially desirable responding in survey research. But is this really a problem around the world? In which countries is this bias strongest, and in which countries can it be ignored?
- There is growing interest in issues related to consumer well-being, as well as a growing realization that transformative consumer research can make a difference around the world (Mick 2005). What are the key drivers of consumer well-being, is their effect moderated by people's cultural and socioeconomic context, and are there systematic and predictable differences in consumer well-being across countries?
- Novak, Hoffman, and Yung (2000, 39) have urged consumer researchers to evaluate "Web sites in terms of the extent to which they deliver these two types [i.e., utilitarian and emotional] of experience." Given the global reach of the Internet and its great influence on

*Martijn G. de Jong is assistant professor of marketing, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands (MJong@rsm.nl). Jan-Benedict E. M. Steenkamp is C. Knox Massey Distinguished Professor of Marketing and Marketing Area Chair, University of North Carolina at Chapel Hill, NC 27599-3490 (JBS@unc.edu). Jean-Paul Fox is assistant professor, Department of Research Methodology, Measurement and Data Analysis, Twente University, P.O. Box 217, 7500 AE Enschede, the Netherlands (G.J.A.Fox@edte.utwente.nl). This article is based on the first author's doctoral dissertation, written when he was a PhD student at Tilburg University. The authors thank AiMark for providing the data and gratefully acknowledge financial support from the Flemish Science Foundation (grant G.0116.04). They also acknowledge the helpful input of the editor, the associate editor, and four reviewers.

John Deighton served as editor and Tulin Erdem served as associate editor for this article.

Electronically published June 21, 2007

consumer behavior, we need to understand these consumption experiences better. Are there universals here? Or is the importance that consumers attach to experiential consumption a "luxury" of industrialized countries?

- Brands are important conduits through which cultural meanings are transferred to individuals (McCracken 1986). Three important brand-related meanings are quality, social responsibility, and prestige (Batra et al. 2000; Roth 1995). Does their importance vary across cultures? Cultural theory would alternatively suggest that prestige connotations are more important in countries with a high degree of power distance, social image meanings are more important in "feminine" countries, and quality associations are more important in individualistic countries.
- Researchers have noted that the construct of *guanxi* plays an important role in social relations in China (Steenkamp 2005). Is this construct unique to China, or does it play a similar role in other collectivistic countries, and perhaps even individualistic countries? How can we integrate such constructs in our theories of consumer behavior?

These issues all have in common that they involve data collection in multiple countries, which requires that the measurement instruments be cross-nationally invariant (Durvasula et al. 1993; Netemeyer, Durvasula, and Lichtenstein 1991; Steenkamp and Baumgartner 1998). 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, 117). The generally accepted view is that, if evidence supporting a measure's invariance is lacking, conclusions based on a research instrument are at best ambiguous and at worst erroneous (Horn 1991). The multigroup confirmatory factor analysis (CFA) model is the dominant approach to investigate cross-national measurement invariance, both in consumer research (Steenkamp and Baumgartner 1998) and other social sciences (Byrne, Shavelson, and Muthén 1989; Vandenberg and Lance 2000).

Despite the advances in cross-national invariance testing using multigroup CFA, two key issues remain unresolved. First, consumer researchers often use five- and seven-point ordinal Likert items to measure latent constructs and the number of scale points may affect reliability and validity (Weathers, Sharma, and Niedrich 2005). However, the multigroup CFA model completely ignores the ordinal nature of the Likert rating scales, which may lead to invalid conclusions regarding measurement invariance (Lubke and Muthén 2004). Measurement invariance may be either over- or understated, thus threatening the validity of cross-national comparisons in consumer research. These results provide further evidence that ordinal data modeling should receive more attention in consumer research (MacKenzie 2003).

Second, the multigroup CFA model requires at least partial invariance, because at least two items exhibit invariance across all countries to make valid cross-country comparisons (Steenkamp and Baumgartner 1998). It is not at all guaranteed that at least two items are invariant, and this constraint becomes ever more problematic the larger the number of countries in one's study (Baumgartner 2004).

The purpose of the present article is to introduce a new cross-national measurement model that addresses both limitations of multigroup CFA. The model is based on item response theory (IRT; Lord and Novick 1968; Samejima 1969). Our model recognizes the ordinal nature of the rating scale, incorporates scale usage, and allows for fully noninvariant item parameters across countries. Although our model allows assessment of measurement invariance for diagnostic purposes, measurement invariance is not needed to make meaningful cross-national comparisons.

The remainder of the article is as follows. First, we review the cross-national measurement invariance literature based on CFA. Next, we introduce our IRT model. Subsequently, we conduct a simulation study to assess the ability of the model to recover its parameter estimates as well as country means and variances. Then, we provide an empirical application of our model, involving an important consumer behavior construct, namely, consumer susceptibility to normative influence (SNI; Bearden, Netemeyer, and Teel 1989), using samples from 11 countries on four continents. We compare the results with the results obtained with multigroup CFA and show that the latter leads to erroneous substantive conclusions. Finally, we present conclusions, limitations, and issues for future research.

MULTIGROUP CFA MODEL

For comparability with the IRT specification, we assume a single construct. Consistent with usual applications of the multigroup CFA model, and without loss of generality, we assume that the number of items is equal across countries. See Baumgartner and Steenkamp (1998) for an extension of the multigroup CFA model that accommodates varying numbers of items across countries. In the CFA model, the relationship between an observed variable and a latent construct is modeled as

$$x_{ik}^g = \tau_k^g + \lambda_k^g \xi_i^g + \delta_{ik}^g, \quad \textbf{(1)}$$

where x_{ik}^g is the observed response to item k ($k = 1, \ldots, K$) for respondent i in country g (with $i = 1, \ldots, N_g$ and $g = 1, \ldots, G$); λ_k^g is the slope (or "factor loading") of the regression of x_{ik}^g on the value of latent construct for respondent i in country g, ξ_i^g; and τ_k^g indicates the expected value of x_{ik}^g when $\xi_i^g = 0$ (Steenkamp and Baumgartner 1998). The model can also be written as $x_i^g = \tau^g + \Lambda^g \xi_i^g + \delta_i^g$, where x_i^g is a $K \times 1$ vector of observed variables in country g, δ_i^g is a $K \times 1$ vector of errors of measurement, τ^g is a $K \times 1$ vector of item intercepts, and Λ^g is a $K \times 1$ vector of factor loadings. Assuming that the measurement errors have zero means, the expectation of x_i^g can be written as $E(x_i^g) = \tau^g + \Lambda^g \kappa^g$, where κ^g is the latent mean of the construct. The variance-covariance matrix among the observed variables x_i^g can be expressed as $V(x_i^g) = \Sigma^g = \Lambda^g \Phi^g \Lambda^{g\prime} + \Theta^g$. In this formula, Φ^g is the variance of the latent

construct and Θ^g is the (usually diagonal) matrix of measurement error variances.

To identify the multigroup CFA model, two constraints are necessary (Steenkamp and Baumgartner 1998). First, it is necessary to assign a unit of measurement to the latent construct. Although there are various ways to do this, the most common approach is to constrain the factor loading of one item (referred to as the marker item) to unity in all countries. Only items that have the same factor loading across countries (i.e., are metrically invariant) may be selected as the marker item. Second, the origin of the scale needs to be identified. Usually, researchers fix the intercept of a latent variable's marker item to zero in each country, so that the mean of the latent variable is equated to the mean of its marker variable. Alternatively, researchers can fix the latent mean at zero in one country and constrain one intercept per factor to be invariant across countries. This item should have invariant factor loadings across countries, which can be checked using empirical criteria such as modification indices and expected parameter changes.

Levels of Invariance

Several tests of cross-national measurement invariance are performed as a prerequisite to conducting comparisons across countries. These tests are necessary in CFA because valid cross-country comparisons require that the scale of the latent variable be the same across countries. Steenkamp and Baumgartner (1998) recommend the use of hierarchical nested models in which the fit statistics of an unconstrained invariance model are examined and compared with the fit statistics of a constrained invariance model by means of a chi-square difference test, which is a likelihood ratio test. Apart from standard chi-square difference tests, the use of fit indexes such as Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) is recommended. The type of invariance in CFA models that is required generally depends on the goals of the study (Steenkamp and Baumgartner 1998). Configural invariance is necessary when the goal is to explore the basic structure of the construct across cultures. Configural invariance is supported if the specified model fits the data well and if all factor loadings are significantly and substantially different from zero.

Metric invariance provides a stronger test of invariance by introducing the concept of equal metrics or scale intervals across countries. 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. Metric invariance (equality of factor loadings) of at least two items is required to compare structural relationships between constructs (Byrne et al. 1989; Steenkamp and Baumgartner 1998). Although one formally only needs one invariant item, an additional invariant item is necessary because of exact identification in case of a single invariant item (any change of metric in the factor loading can be compensated for by change in the metric of the latent construct). To test the item's invariance, an overidentified model is necessary with another invariant item.

Consumer researchers are often interested in comparing the means on the construct across countries. In order for such comparisons to be meaningful, scalar invariance (equality of intercepts) of the items is required (Meredith 1993). Scalar invariance addresses the question 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 of at least two items that also exhibit metric invariance is necessary to conduct valid cross-national comparisons in construct means (Steenkamp and Baumgartner 1998) for the same reason as for metric invariance.

Limitations of CFA

The multigroup CFA framework has several important limitations. First, testing for partial invariance is generally an exploratory post hoc method, subject to capitalization on chance. MacCallum, Roznowski, and Necowitz (1992) recommend that 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. In addition, if there are few invariant items, the usual tests for differential item functioning may identify an invariant item as being noninvariant due to the fact that the model also tries to fit the other noninvariant items (Holland and Wainer 1993).

Second, to make substantive comparisons, at least two items should exhibit invariance across countries. This requirement is independent of scale length. But when the measurement instrument consists of only few items, or when the number of countries increases, this requirement is likely to be problematic (Baumgartner 2004). When measurement invariance is not satisfied, subgroups of countries have to be found that are measurement invariant (Welkenhuysen-Gijbels, Billiet, and Cambré 2003). However, researchers usually want to compare all countries.

Third, multigroup CFA does not recognize the ordinal nature of the rating scale. Recent simulation studies have shown that ignoring the ordinal nature of the data is problematic in multigroup research (Lubke and Muthén 2004). The CFA methodology assumes that the observed data are multivariate normally distributed, and, therefore, tests of measurement invariance focus on the regression intercepts τ_k^g and factor loadings λ_k^g. However, the set of parameters required to achieve measurement invariance across countries is different for ordinal data. Although there are multiple ways to conceptualize ordinal data, a common data-generating mechanism starts with an unobserved continuous outcome and states that a response category is chosen above a

lower category if the continuous latent variable exceeds a certain threshold. These thresholds are not modeled in CFA. As a result, measurement invariance tests based on the CFA methodology can indicate that measurement invariance is satisfied, when it is not, and vice versa, complicating cross-national comparisons of the latent construct (Lubke and Muthén 2004). However, these thresholds can be modeled by IRT models for polytomous (ordinal) data.

IRT MODEL

Below, we describe the IRT approach. We start with an overview of the general aspects of IRT for polytomous data. Although IRT models have been popular for dichotomous items, Samejima (1969, 1972) extended IRT models to polytomous items with multiple ordered response categories. Next, we discuss the traditional multigroup IRT model and how the different countries can be linked together so that the latent variable is measured on the same scale across countries. Like CFA, previous multigroup IRT models require certain levels of invariance to allow for valid country comparisons (May 2005; Meade and Lautenschlager 2004).

Subsequently, our new IRT model is introduced. Our model takes not only mean differences into account (like Holland and Wainer 1993) but also scale-usage differences. Moreover, our model does not require cross-national measurement invariance for valid country comparisons. Nevertheless, invariance tests may be useful for diagnostic purposes—for example, to better understand response behavior in different countries (see Wong et al. 2003). Hence, we conclude this section with a discussion on invariance tests in the context of our IRT model.

IRT for Ordinal Response Data

IRT models posit a reflective (see Jarvis, MacKenzie, and Podsakoff 2003), nonlinear relationship between an underlying latent construct and the observed score at the item level. Despite many advantages over the classical test theory paradigm, IRT models have been conspicuously absent from the marketing literature (see Balasubramanian and Kamakura [1989]; Bechtel [1985]; and Singh, Howell, and Rhoads [1990] for exceptions).

IRT has mainly been used in marketing for adaptive surveys, that is, surveys in which questions are adapted based on an individual's previous responses (see Balasubramanian and Kamakura [1989] for an example of the tailored interview process). IRT models for ordinal data are conceptually somewhat similar to ordinal/limited dependent data models in the econometrics literature (Franses and Paap 2001; Greene 2003; Maddala 1983). However, in IRT models, there are multiple ordinal items that reflect a latent construct, while for the ordinal data models in econometrics, there is usually a single ordinal variable.

The item response function (IRF) is the nonlinear monotonic function that accounts for the relationship between a respondent's value for latent variable ξ_i^g and the probability of a particular response on an item. Local independence is assumed; that is, there is no relationship between the respondent's item responses given ξ_i^g. Polytomous IRT models deal with responses to items measured on C ordered response categories. For example, the five-point Likert item commonly used in marketing research has $C = 5$ ordered response options, such as "Strongly disagree," "Disagree," "Neither agree nor disagree," "Agree," and "Strongly agree." In a cross-national setting with G countries, the graded response model for country g is given by

$$\begin{aligned} P(x_{ik}^g = c|\xi_i^g, a_k^g, \gamma_{k,c}^g, \gamma_{k,c-1}^g) &= \Phi(a_k^g\xi_i^g - \gamma_{k,c-1}^g) - \Phi(a_k^g\xi_i^g - \gamma_{k,c}^g) \\ &= IRF_{k,c-1}^g - IRF_{k,c}^g, \end{aligned} \quad \textbf{(2)}$$

where $\Phi(.)$ is the standard normal cumulative distribution function. This model specifies the conditional probability of a person i in country g, responding in a category c ($c = 1, \ldots, C$) for item k, as the probability of responding above $c - 1$, minus the probability of responding above c. The parameter a_k^g is called the discrimination parameter for item k in country g and is conceptually similar to the factor loading λ_k^g in the CFA setting, in that it represents the strength of the relationship between the latent variable and item responses (Reise, Widaman, and Pugh 1993). Useful items have a large discrimination parameter.

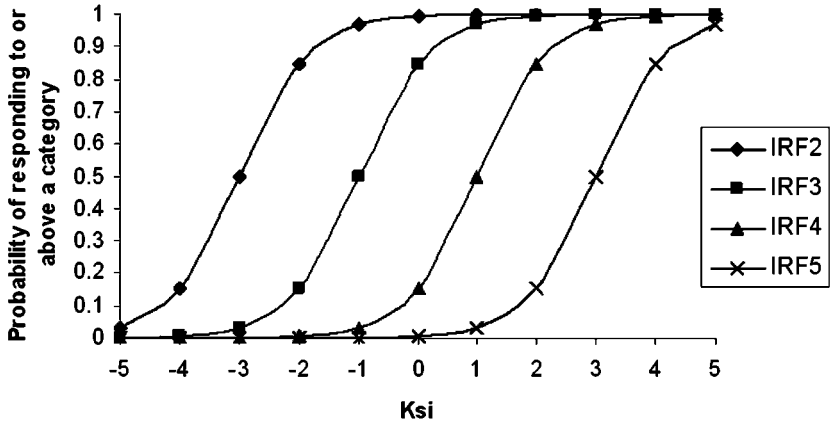
The thresholds $\gamma_{k,c}^g$ are measured on the same scale as ξ_i^g and determine the difficulty of responding above a certain response category c. The threshold $\gamma_{k,c}^g$ is defined as the value on the ξ_i^g scale so that the probability of responding above a value c is .5, for $c = 1, \ldots, C - 1$. In equation 2, one can put $\gamma_{k,0}^g = -\infty$, $\gamma_{k,C}^g = \infty$, so that only the thresholds for the categories 1 through $C - 1$ need to be considered. For illustration, we draw the IRFs for an item k on a five-point Likert scale in country g with $a_k^g = 1$, $\gamma_{k1}^g = -3$, $\gamma_{k2}^g = -1$, $\gamma_{k3}^g = 1$, $\gamma_{k4}^g = 3$ in figure 1.

The IRF curves display the probability of responding above a certain rating scale point as a function of a person's position on the underlying latent construct. Only four curves are shown, as, by definition, the probability of responding above $c = 5$ is zero. For instance, IRF_2 graphs the probability of responding above $c = 2$ for varying levels of ξ_i^g. Suppose that a respondent has $\xi_i^g = -2$; s/he then has a probability of .85 of responding above $c = 1$, a probability of .15 of scoring above $c = 2$, and a probability of almost zero of responding above $c = 3, 4, 5$. Thus, $c = 2$ is the most likely outcome.

The IRFs, displayed in figure 1, can be used to compute the probability of a category response by equation 2. The category response functions (CRF) for the item with the item parameters given above are displayed in figure 2. Note that the values for γ correspond to the intersection of two successive CRFs. For instance, for $\xi_i^g = \gamma_{k1}^g = -3$, the CRFs for categories 1 and 2 intersect. Further, it can be seen that a respondent with $\xi_i^g = -2$ has a probability of .15 of responding $c = 1$, a probability of .69 of responding $c = 2$, a probability of .15 to answer $c = 3$, a probability of .01

FIGURE 1

ILLUSTRATIVE ITEM RESPONSE FUNCTIONS



of responding $c = 4$, and a probability of zero of responding $c = 5$. Across all categories, the response probabilities within respondents sum to one.

Cross-National Differences in Scale Usage

An important advantage of using IRT is that the ordinal nature of the rating scale and, thus, rating scale usage (Rossi, Gilula, and Allenby 2001) are taken into account. Indeed, it has been shown that countries differ in rating scale usage, such as extreme responding and yea-saying, and that this may seriously bias one's substantive findings (Baumgartner and Steenkamp 2001). To illustrate how IRT accounts for scale usage, consider a country where respondents are reluctant to use the ends of the rating scale for a particular item *k*. In this case, the outer category thresholds would be larger in absolute sense, increasing the probability of middle responses, while simultaneously reducing the odds of an extreme response. This process is illustrated in figure 3, where we set $a_k^g = 1, \gamma_{k1}^g = -5, \gamma_{k2}^g = -1, \gamma_{k3}^g = 1, \gamma_{k4}^g = 5$.

Comparing figures 2 and 3, it can be seen that, for the same values of ξ_i^g, the probability of responding in categories 2 or 4 becomes larger, while the odds of responding in categories 1 and 5 are very small. So, although the latent score is the same, scale usage in a country on item *k* determines the response on the rating scale. Analogously, if a country rates highly on yea-saying on a particular item, the thresholds for categories 4 and 5 become smaller. We note that the bias should not be completely uniform across items (Thissen, Steinberg, and Gerrard 1986). Recent evidence indeed shows that the bias is different for different items (De Jong et al. 2007).

Identification and Linking Groups

As in the CFA models, two issues need to be addressed. First, the IRT model needs identification restrictions, since

FIGURE 2

ILLUSTRATIVE CATEGORY REPONSE FUNCTIONS

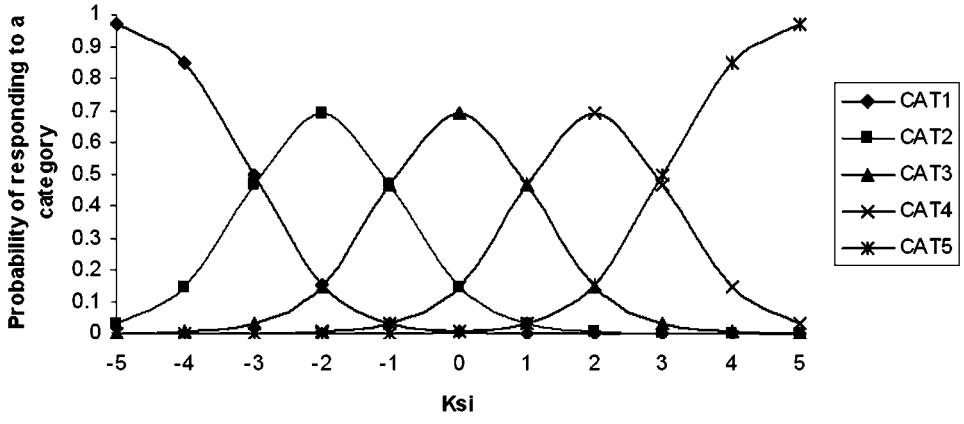
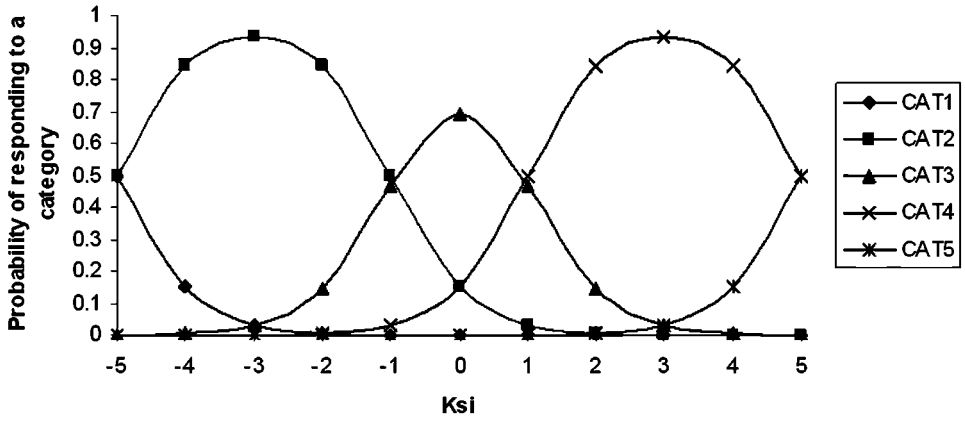


FIGURE 3

CRFS FOR COUNTRY WITH LOW TENDENCY TO EXHIBIT EXTREME RESPONDING



the latent variable has no definite origin. Second, we specified a separate IRT model for each country *g*, without linking the *G* models. To make meaningful substantive comparisons across countries, the IRT models should be linked to ensure that the numerical values for the latent variable across countries are on the same measurement scale. If the scores on the latent variables are not on the same scale, differences between countries in mean levels or in structural relations of the construct with other constructs might be spurious.

To scale the latent variable, single-group IRT models usually specify a distribution for the latent variable with mean zero and variance one. It is also possible to use item parameter restrictions to fix the scale of the latent variable. In cross-national settings, mere standardization in each country without linking the countries renders item parameters incomparable across groups. An approach that has been commonly used in previous research is fixing the mean to zero and variance to one in the reference group, freely estimating the mean in the other groups, while fixing the variance in the other groups to some value determined by a trial and error analysis (Reise et al. 1993). Thus, the variance of the latent variable is not estimated freely across groups.

If no further restrictions are employed, and all items are estimated freely across countries, the model is identified, but the metric for ξ is not common across countries. Therefore, additional restrictions are necessary to link the groups. Multigroup IRT models to date impose invariance restrictions on the item parameters (May 2006; Meade and Lautenschlager 2004; Reise et al. 1993) to make the scale common across countries. A minimum identifying constraint is that, for at least one anchor item, the item parameters are invariant across countries. In that case, calibrating the rest of the items together with the anchor item results in a common scale for ξ across countries. Note that this still requires an item that is known (or assumed) to be fully invariant across countries.

Hierarchical IRT

We propose a new approach to identify and link groups. We first model differential item functioning, including scale usage differences across countries using a random-effects ANOVA formulation. Random-effects IRT specifications for binary response data that allow for random item variation were proposed by Janssen et al. (2000). However, in their article, the grouping was based on items, rather than on countries. In addition, the data in our setting are polytomous. We model random item parameter variation as:

$$\gamma_{k,c}^{g} = \gamma_{k,c} + e_{k,c}^{g},\ e_{k,c}^{g} \sim N(0, \sigma_{\gamma_k}^{2}) \quad \text{for } c = 1, \ldots, C-1,\ \gamma_{k,1}^{g} \leq \ldots \leq \gamma_{k,C-1}^{g} \tag{3}$$

$$a_k^g = a_k + r_k^g,\ r_k^g \sim N\left(0, \sigma_{a_k}^2\right),\ a_k^g \in (0, A],\ a_k \in (0, A]. \tag{4}$$

Equation 3 implies that each scale threshold $\gamma_{k,c}^{g}$ for a particular item k in country g is modeled as an overall mean threshold $\gamma_{k,c}$, plus a country-specific deviation $e_{k,c}^{g}$. Analogously, equation 4 posits that the discrimination parameter a_k^g is the sum of an overall mean discrimination parameter and country-specific deviation. The prior specification imposes that the discrimination parameter is positive and smaller than some positive number *A*. The variances of the threshold and discrimination parameters are allowed to vary across items. In our model, there is no longer a need to classify items as being invariant or noninvariant.

When calibrating the item parameters, it is important to model the heterogeneity in the latent variable. Thus, a hierarchical structure is imposed on ξ_i^g by letting

$$\xi_i^g = \xi^g + v_i^g,\ v_i^g \sim N(0, \sigma_g^2), \tag{5}$$

$$\xi^g \sim N(\xi, \tau^2). \quad \textbf{(6)}$$

In other words, the position on the latent scale for respondent i in country g is sampled from the country average ξ^g with variance σ_g^2. The country average is drawn from a distribution with average ξ and variance τ^2. This random-effects approach for the latent variable is consistent with recent work on multilevel latent variable modeling in psychometrics (Fox and Glas 2001, 2003). We note that equations 5 and 6 can be extended by incorporating individual and country-level covariates, but this is not pursued here.

When the random-effects structure for item parameters is combined with the random-effects structure for the latent variable, there is an identification problem. Each country mean can be shifted by changing the country mean, ξ^g, as well as by uniformly shifting the country-specific threshold values, $\gamma_{k,c}^g \quad \forall\, k$. We fix the mean of country g by restricting the country-specific threshold parameters in such a way that a common shift of these threshold values is not possible. This can be done by setting $\sum_k \gamma_{k,3}^g = 0$. Since this restriction is applied in each country, the mean of the metric of the latent variable is identified via restrictions on the country-specific threshold parameters.

Analogously, the country variances can be shifted both by σ_g^2, as well as by uniform changes in the discrimination parameters (i.e., setting $a_{k,new}^g = a_k^g \times d \quad \forall\, k$). To fix the country-specific variances, we need to impose a restriction that a common shift of country-specific discrimination parameters is not possible, which can be done by imposing the condition that, across items, the product of the discrimination parameters equals one in each country g ($\prod_k a_k^g = 1 \quad \forall\, g$). Hence, both the mean and variance of the latent variable in each country are fixed, and the scale remains common due to the simultaneous calibration of the multilevel structures for item parameters and latent variable.

The hierarchical Bayesian framework allows for borrowing of strength across countries. Previous multigroup CFA research models country means/variances, factor loadings, and item intercepts as separate parameters, without borrowing strength across countries. The same holds for previous multigroup IRT research (i.e., discrimination, threshold, and country mean and variance are modeled as separate parameters). By borrowing strength, we can place less restrictive assumptions on measurement invariance, while retaining the possibility to let the various parameters fluctuate across countries. In table 1, we present an overview table to contrast our specification with previous multigroup IRT and CFA models.

IRT Estimation

Both marginal maximum likelihood techniques and Bayesian techniques have been used in previous multigroup IRT research (e.g., Bolt et al. 2004; May 2006; Meade and Lautenschlager 2004; Reise et al. 1993; Thissen, Steinberg, and Wainer 1988, 1993). We use Bayesian techniques to estimate the model parameters. The Bayesian approach requires the specification of a full probability model. To obtain draws from the posterior distribution, we use a data-augmented Gibbs sampler (Tanner and Wong 1987) with a Metropolis-Hastings step for the threshold parameters. Estimation details, including the priors, are described in appendix A.

IRT-Based Invariance Testing

Although our hierarchical IRT model does not require invariance across countries to make substantive comparisons, we describe the various levels of invariance that can be imposed on the IRT model below. These tests of invariance would mainly serve as a diagnostic tool, for example, to see whether or not items are culturally biased or to investigate other aspects of either the measurement or the structural model (e.g., Raju, Byrne, and Laffitte 2002; Reise et al. 1993; Wong et al. 2003). Previous research has considered only invariance of the discrimination parameters and the threshold parameters and not the invariance of the latent variable variance because it could not be freely estimated (see Bolt et al. 2004; Meade and Lautenschlager 2004; Reise et al. 1993). Our model also allows tests of factor variance invariance, that is, invariance of the latent variable variance across countries. Full-item parameter invariance is satisfied if for all items k: $a_k^1 = a_k^2 = \ldots = a_k^G$, $\gamma_{k1}^1 = \gamma_{k1}^2 = \ldots = \gamma_{k1}^G$, $\gamma_{k2}^1 = \gamma_{k2}^2 = \ldots = \gamma_{k2}^G$, $\ldots$, $\gamma_{kC-1}^1 = \gamma_{kC-1}^2 = \ldots = \gamma_{kC-1}^G$.

We assess item parameter invariance via Bayes factors (Kass and Raftery 1995; Newton and Raftery 1994). The proposed model, M_1, with varying item parameters is compared to a model, M_2, with fixed item parameters across countries. The Bayes factor is the ratio of the two marginal likelihoods, the marginal likelihood of the data under model M_1 and M_2. Large values of the Bayes factor BF_{12} indicate

TABLE 1

OVERVIEW OF MULTIGROUP LATENT VARIABLE MODELS

	Latent variable heterogeneity (separate country means and variances)	Random effects structure for item parameters	Invariance requirements on items
Previous multigroup IRT approaches	Yes (separate means and variances)	No	Yes
Multigroup CFA approach	Yes (separate means and variances)	No	Yes
This article	Yes (random-effects structure)	Yes	No

a preference for model M_1. The Bayes factors are computed via importance sampling (Newton and Raftery 1994). Bayesian inferences regarding the variances of the item parameters are based on their marginal posterior distributions. Factor variance invariance is tested, using a Bayesian parallel to Bartlett's test of equal variances, while means can be compared using a Bayesian ANOVA. We refer to appendix B for more details.

SIMULATION STUDY

The purpose of the simulation study was to examine (1) whether the country-specific discrimination and threshold parameters can be recovered and (2) whether the country-specific latent means and variances can be recovered, under (3) the condition that no measurement invariance constraints are imposed on the model. For this purpose, we generate a data set with no cross-nationally invariant items. That is, there is variation in the values of the item parameters for each item across countries. The multigroup CFA approach would not be feasible in this case, because metric invariance is not satisfied for any item. In addition, mean comparisons would not be possible due to differences in scale usage for all items. However, as shown below, the IRT model does allow researchers to conduct substantive cross-national comparisons, even though measurement invariance is not fulfilled, because all respondents in all countries are calibrated on the same latent scale.

Data were generated according to the random effects specifications in equations 3–6 with 10 countries, 1,000 respondents per country. There are 10 items, and each item is measured on a four-point Likert scale. In the simulation design, both the discrimination parameters and the threshold parameters are generated so that they vary randomly across nations. For the threshold parameters, the standard deviations range from 0.45 to 0.65 across items k, while for the discrimination parameters, standard deviations range from 0.15 to 0.40 across items k.

For the item parameters, we present scatter plots of estimated versus true parameters in figure 4. The true values are accurately recovered by the model. This applies to both the discrimination parameters and the threshold parameters. Regressing the estimated discrimination parameters on the true discrimination parameters results in a regression slope of 0.97, where the 95% confidence interval includes one, and an R^2 of 0.91. Similarly, a regression of estimated threshold parameters on true threshold parameters yields a regression slope of 0.99, with a 95% confidence interval that includes one, and an R^2 of 0.99.

In table 2, we present the true versus the estimated country means and variances. As can be seen, parameter recovery is accurate. On average, estimated country means differ only 1.33% from the true country means, while the difference in the country latent score variances is 5.68% due to sampling error. Thus, the simulation study revealed that our model was able to accurately recover country-specific means, variances, discrimination, and threshold parameters, although there was not a single invariant item.

APPLICATION TO CONSUMER SUSCEPTIBILITY TO NORMATIVE INFLUENCE

Consumer Susceptibility to Normative Influence

We apply our model to real cross-national data on consumers' susceptibility to normative influence. Consumers differ in the degree to which they are influenced in their attitudes and behavior by the norms of the social system, that is, in their susceptibility to normative influence (SNI). The consumer behavior literature recognizes that individuals' behavior cannot be fully understood unless consideration is given to the effect of a person's SNI on development of attitudes, aspirations, and behavior (Bearden et al. 1989). SNI has been linked to various aspects of consumer behavior such as attitudes toward brands (Batra et al. 2000), advertising (Mangleburg and Bristol 1998), and consumption alternatives resulting from globalization (Alden, Steenkamp,

FIGURE 4

ESTIMATED VERSUS TRUE ITEM PARAMETERS

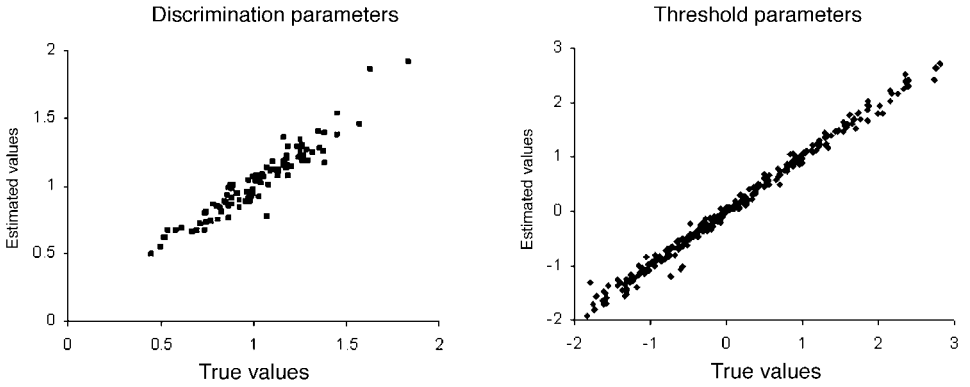


TABLE 2

TRUE AND ESTIMATED COUNTRY MEANS AND VARIANCES

	Country mean		Country variance	
Country	True value	Estimated value	True value	Estimated value
1	−.299	−.323	.332	.314
2	2.399	2.274	.645	.510
3	.154	.161	1.212	1.230
4	−.823	−.800	.889	.870
5	−.273	−.241	.461	.484
6	−.121	−.131	.593	.598
7	.271	.276	.466	.461
8	−.862	−.812	1.467	1.234
9	.321	.303	1.381	1.329
10	1.688	1.607	1.910	1.626

and Batra 2006), consumer confidence (Bearden, Netemeyer, and Teel 1990), protective self-presentation efforts (Wooten and Reed 2004), purchase of new products (Steenkamp and Gielens 2003), and consumer boycotts (Sen, Gürhan-Canli, and Morwitz 2001), among others. Consumers who rank high on SNI tend to rank lower on self-esteem and higher on motivation to comply with the expectations of others, interpersonal orientation, and attention to social comparison information (Bearden et al. 1989, 1990). Most SNI research has been carried out in the United States, despite the obvious importance of normative influences in other, for example, collectivistic cultures (Kagitcibasi 1997).

Consumers in some countries may on average rank higher on SNI than consumers in other countries, due to systematic differences in the national cultural environment. Culture is a powerful force shaping people's perceptions, dispositions, and behaviors (Triandis 1989) and is reflected in "persistent preferences for specific social processes over others" (Tse et al. 1988, 82). We expect that national-cultural individualism is especially important for understanding cross-national differences in SNI. National-cultural individualism pertains to the degree to which people in a country prefer to act as individuals rather than as members of a group. Collectivistic cultures are conformity oriented and show a higher degree of group behavior and concern to promote their continued existence.

The conformity pressure and the close-knit social structure will also result in less divergence in attitudes compared with individualistic countries because divergence in attitudes is less valued in collectivistic cultures (Kagitcibasi 1997). In individualistic societies, the social fabric and group norms are much looser. People tend not to follow social norms but rather make decisions and initiate behaviors independently of others (Roth 1995). A child already learns very early to think of herself as "I" instead of as part of "we," while the converse holds for collectivistic societies (Hofstede 2001). Thus, consumer cultural theory suggests that consumers living in individualistic countries (1) rank on average lower on SNI and (2) exhibit more divergence in their SNI attitudes compared with consumers living in collectivistic countries.

Method

The data collection was part of a large global study on consumer attitudes. Data collection was carried out by two global marketing research agencies, GfK and Taylor Nelson Sofres. The total sample for the present application comprises 5,484 respondents in 11 countries from four continents, namely, Brazil, China, France, Japan, the Netherlands, Poland, Russia, Spain, Taiwan, Thailand, and the United States. The number of respondents per country ranges from 396 (Taiwan, Russia) to 546 (Spain). Given the importance of the United States, the marketing research agencies decided to put an additional effort in sampling respondents from the United States. Therefore, the number of respondents for the United States is 1,181. The samples in each country were drawn so as to be broadly representative of the total population in terms of region, age, education, and gender.

For the United States, France, Spain, Japan, and the Netherlands, a Web survey was used in which respondents in script panels of GfK and Taylor Nelson Sofres were invited to participate in the project by an e-mail in the local language. The e-mail contained a short description, a hyperlink to go to the survey, and an estimate of the time needed to complete the survey. At the end of the fieldwork period, respondents were paid by the local subsidiary of the global marketing research agencies.

For China and Russia, Internet surveys were administered using mall intercepts. For the mall intercepts, the first step was to select multiple regions/locations for the fieldwork. Next, a space was rented that had an Internet connection for two to five PCs or laptops (e.g., Internet cafes, subsidiaries of offices, test halls for product tests) and offered the possibility to intercept appropriate shoppers/respondents walking in the street, using street recruiters.

Finally, in Brazil, Taiwan, and Thailand, a hard-copy survey instrument was used, which was also administered in mall intercepts. The hard-copy tool was designed so that the layout was exactly the same as in the Internet survey. The staff for the hard-copy mall intercepts generally consisted

TABLE 3

SNI ITEMS

Item	Description
1	If I want to be like someone, I often try to buy the same brands that they buy.
2	It is important that others like the products and brands I buy.
3	I rarely purchase the latest fashion styles until I am sure my friends approve of them.
4	I often identify with other people by purchasing the same products and brands they purchase.
5	When buying products, I generally purchase those brands that I think others will approve of.
6	I like to know what brands and products make good impressions on others.
7	If other people can see me using a product, I often purchase the brand they expect me to buy.
8	I achieve a sense of belonging by purchasing the same products and brands that others purchase.

of a field supervisor, responsible for answering respondents' questions and monitoring the whole fieldwork, a logical controller, responsible for logical control and sampling quotas, and three to four street recruiters.

SNI was measured using the eight-item scale developed by Bearden et al. (1989). This unidimensional scale has been extensively validated and is the most frequently used instrument to measure SNI. The items are listed in table 3. The SNI items were translated into all local languages by professional agencies. Next, the translated items were translated back into English, using native speakers from the local countries. In each survey, modifications were made based on discussions between the translators, one of the authors, and the headquarters of the marketing research agencies to maintain consistency in changes across all countries. All items were measured on a five-point Likert scale.

We randomly dispersed the items throughout the questionnaire. There is a debate in the literature whether items pertaining to the same construct should be randomized in the questionnaire or grouped together (Bradlow and Fitzsimons 2001). The idea behind randomization is to hide the purpose of the instrument from the respondent, thus reducing response biases such as a desire to look good to others (e.g., evaluation apprehension) and to oneself (cognitive consistency and ego defense mechanisms). But randomization may also reduce reliability (Bradlow and Fitzsimons 2001). However, low reliability was not an issue in our study, as, in all countries, the reliability of SNI exceeded the .70 cutoff.

We used Bayesian routines programmed in Fortran for the IRT model. The Bayesian routines are linked to S-Plus®. The software to estimate this model can be obtained from the authors. Other researchers can easily estimate their own models by adapting the number of items and countries. The number of countries groups matters for the choice of a fixed versus random effects approach.

IRT Results

We estimate our hierarchical IRT model using Markov Chain Monte Carlo (MCMC) techniques. Convergence of the chains is checked using the CODA software (Best, Cowles, and Vines 1995), which contains multiple standard convergence diagnostics. Multiple chains for different starting values were run, and convergence occurred quickly for most parameters. We ran multiple chains for different starting values. The first 10,000 iterations are discarded, and 30,000 posterior draws are subsequently used to estimate the model parameters. We present the results of the model, and in the next section we consider a number of invariance tests for diagnostic purposes.

Table 4 presents the estimation results for the discrimination parameters. The items generally discriminate well for each country, given the posterior distribution of the latent variable. Table 4 further shows that there are substantial cross-national differences in the discrimination power of any specific item.

On average, item 3 has a lower discrimination parameter than the other items, indicating that this item measures the SNI construct somewhat less well than the other items. Interestingly, it is the only item that refers to a specific con-

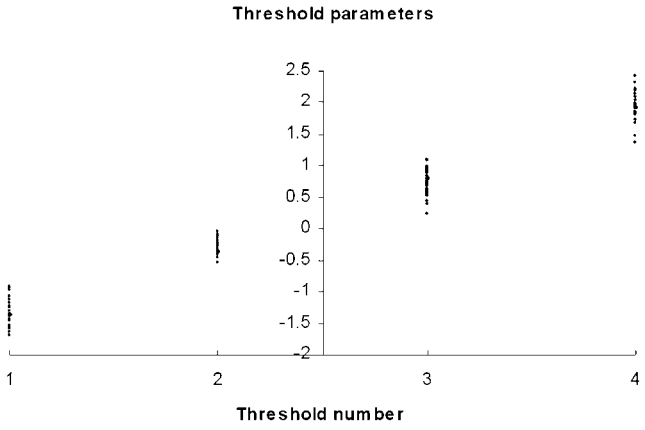
TABLE 4

DISCRIMINATION PARAMETERS FOR SNI SCALE

	Item							
Country	1	2	3	4	5	6	7	8
France	.862	.654	.582	1.260	1.272	1.171	1.526	1.089
Netherlands	.794	.826	.585	1.231	1.328	1.087	1.150	1.292
Spain	.883	.710	.894	.946	1.312	.895	1.265	1.287
China	1.028	1.036	.644	1.299	1.078	.703	1.320	1.152
Poland	.931	.881	.276	1.246	1.420	1.101	1.581	1.512
Brazil	.880	.734	.843	.768	1.478	1.174	1.337	1.063
Thailand	.845	.814	.524	1.309	1.244	1.033	1.232	1.381
Russia	.879	1.227	.403	1.106	1.413	1.265	.855	1.418
United States	.767	.822	.610	1.137	1.319	1.045	1.281	1.306
Taiwan	.897	.914	.581	1.307	1.308	.592	1.563	1.378
Japan	1.040	.781	.724	1.305	1.251	.682	1.258	1.248

FIGURE 5

CROSS-NATIONAL VARIATION OF POSTERIOR MEAN THRESHOLD PARAMETERS FOR SNI ITEM 6



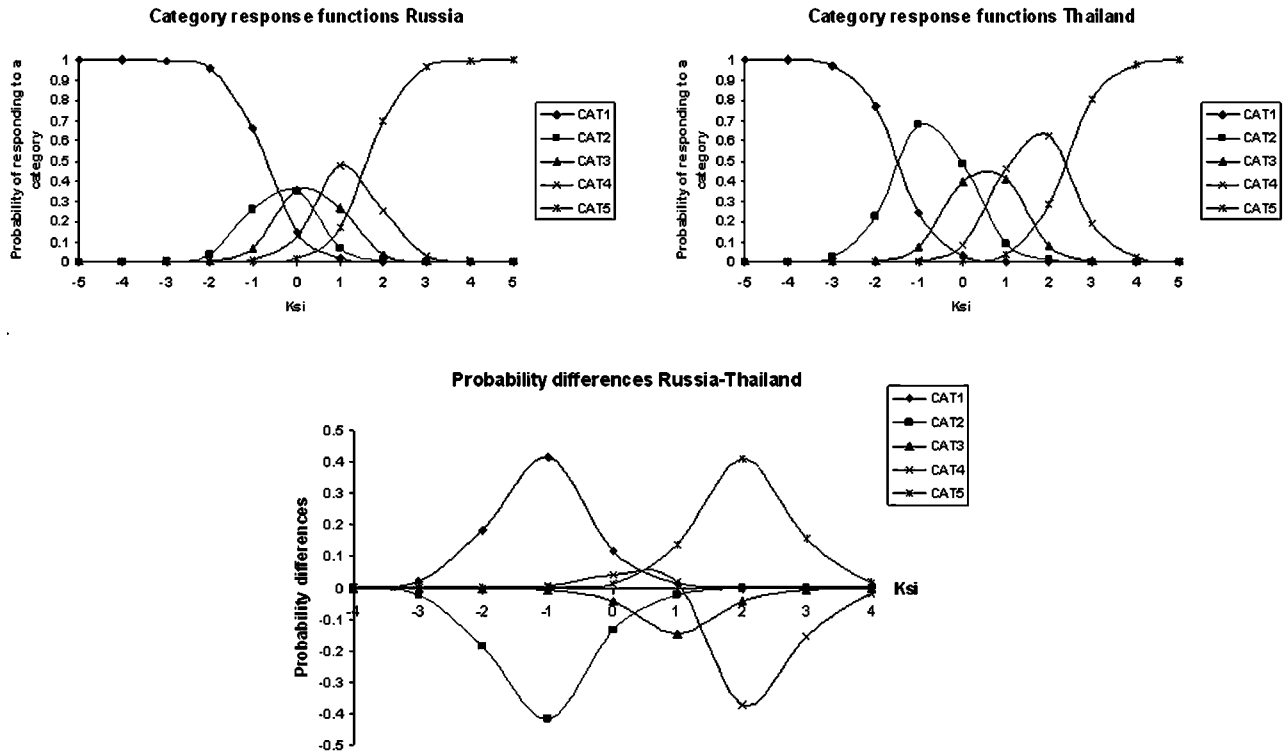
sumption domain (fashion styles). From a scale reliability point of view (although not necessarily from a content/predictive validity standpoint), items 5, 7, and 8 are on average the best items. These results are quite consistent across countries. Thus, if a researcher wants to use a maximally reliable short-form of SNI because eight items is too much (see Burisch 1984), items 5, 7, and 8 would be prime candidates. For the IRT model, it does not matter whether items 5, 7, and 8 are invariant. For CFA, we would need two items in the short scale to be invariant.

Next, we turn to the threshold parameters. Since the items are measured on a five-point scale, there are four thresholds per item. Thus, each country has $4 \times 8 = 32$ threshold parameters, so in total, there are 352 threshold parameters. As was the case for the discrimination parameters, there is substantial cross-national variation. For illustrative purposes, we plot the posterior means of the threshold parameters for item 6 in figure 5.

To illustrate the effect of the threshold parameters on the probability of responding to a certain Likert response category, we plot the CRFs for item 8 for Thailand and Russia in figure 6. Comparing Thailand and Russia, we see that for equal true scores, the probability of responding in categories 1 and 5 (i.e., "Strongly disagree" and "Strongly agree") is much smaller in Thailand than in Russia. Consider for instance a respondent with a moderately low latent SNI score of $\xi = -1$. In Thailand, this respondent has a probability of .24 to choose the response category 1, while in Russia, this probability is .66. Thus, for this item, there is a difference of .42 in the probability between Russian and Thai respondents in checking response category 1 on the five-point scale, even though they hold the same underlying true

FIGURE 6

CRFS FOR ITEM 8



opinion. However, Thai respondents with a latent SNI score of $\xi = -1$ have a vastly greater probability of checking response category 2 than Russian respondents with the same underlying score (.69 versus .28, a difference of .41). A respondent with $\xi = 2$ has a probability of only .31 of choosing the response "Strongly agree" in Thailand, while this probability is .71 in Russia.

Summarizing, there is substantial evidence of differential item functioning across countries. The hierarchical IRT model accommodates these differences and puts the estimates of the latent variable in different groups on the same scale. Furthermore, we can test whether countries differ significantly in their mean SNI score. We do this by computing a Bayesian ANOVA, based on an $F(10, 5, 473)$-statistic (see app. B). Indeed, countries differ in their mean score on SNI ($p < .001$). The Bayesian Bartlett test for factor variance invariance shows that there are also cross-national differences in within-country heterogeneity on SNI ($p < .001$). The heterogeneity in SNI scores is properly modeled by taking the hierarchical structure for the latent variable into account and by allowing for different within-country variances.

IRT-Based Invariance Tests

Although our model does not require invariance of parameters across countries for valid cross-national comparisons, invariance tests are of interest for diagnostic purposes, that is, to better understand response behavior for specific items and countries (e.g., Raju et al. 2002; Reise et al. 1993; Wong et al. 2003). We test the plausibility of several competing models using marginal log-likelihoods and Bayes factors (Kass and Raftery 1995). Three models are considered. The first model (M_1) has invariant discrimination and threshold parameters, that is, $a_k^g = a_k \quad \forall\, g$, and $\gamma_{kc}^g = \gamma_{kc}$, $c = 1, \ldots, C-1 \quad \forall\, g$. The second model ($M_2$) relaxes all invariance constraints on the threshold parameters γ (thus, only the discrimination parameters $\boldsymbol{a}$ are kept invariant across countries), and the third model (M_3), which is the most flexible one and for which the results are reported above, relaxes all invariance constraints (both on $\boldsymbol{a}$ and γ). The marginal log-likelihoods and Bayes factors (assuming $P(M_1) = P(M_2) = P(M_3)$) of the different models versus model M_3 all indicate that the posterior probability of M_3 given the data (Berger and Delampady 1987) is much higher than the probability of models M_1 and M_2. Relaxing the invariance constraints on the discrimination parameters and on the threshold parameters yields a large improvement in fit.

The model comparison results are consistent with the earlier observation that there is substantial variation in the discrimination and threshold parameters. This indicates that a particular item does not perform equally well in different countries (i.e., does not discriminate equally well between respondents in different countries) and that there are substantial cross-national differences in response behavior on the five-point rating scale.

CFA Results

We estimated a one-factor model for the 11 countries, using LISREL. The configural invariance model specifies the same pattern of factor loadings across countries and serves as a baseline model. We choose item 2 as the marker item (based on modification indices, this choice seemed best). In multigroup CFA, one does not know a priori which item can be used as a marker item. The usual procedure is to select one item, impose full metric invariance, and examine the modification indices. If another item has smaller modification indices, the model is reestimated using this item as a marker item (Steenkamp and Baumgartner 1998). The fit of the configural invariance model is good (see table 5). Although the χ^2 was significant—which is not unexpected given the large sample size—other indicators exceeded conventional cutoff levels: $\chi^2(220) = 914.4$ ($p < .001$), RMSEA $= 0.0791$, CFI $= 0.973$, TLI $= 0.962$. The within-country completely standardized loadings are relatively high. Of the $11 \times 8 = 88$ factor loadings, 38 standardized loadings exceed 0.6, and 49 loadings exceed 0.5. Based on these results, we conclude that the SNI scale exhibits configural invariance.

In the next step, we test for full metric invariance by constraining all factor loadings to be equal across countries. The fit of this model deteriorates substantially ($\Delta\chi^2(70) = 343.4$, $p < .001$); RMSEA and TLI, which both take fit and model parsimony into account, deteriorate. In a recent extensive simulation study, Cheung and Rensvold (2002) found that ΔCFI is a particularly robust statistic for testing multigroup invariance constraints and reported that "a value of ΔCFI smaller than or equal to $-.01$ indicates that the null hypothesis of invariance should not be rejected" (251). Since in our application, ΔCFI decreased by 0.019, we conclude that full metric invariance is not supported.

Examination of the modification indices (MIs) revealed that the deterioration in fit was due largely to a lack of invariance of four factor loadings, namely, the loadings of items 3 and 8 in Spain (MI $= 20.8$ and 31.2, respectively) and factor loadings for item 6 in Taiwan and China (MI $=$

TABLE 5

MULTICOUNTRY CFA MODEL COMPARISONS FOR SNI SCALE

	χ^2	*df*	RMSEA	CFI	TLI
Configural invariance	914.4	220	.0791	.952	.933
Metric invariance	1,257.8	290	.0820	.933	.929
Final partial metric invariance	1,100.8	286	.0748	.943	.939
Scalar invariance	2,153.7	352	.101	.875	.891
Final partial scalar invariance	1,279.8	322	.0767	.934	.936
Factor variance invariance	1,369.1	332	.0784	.928	.933
Partial factor variance invariance	1,289.5	327	.0762	.933	.937

NOTE.—RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index.

57.7 and 29.9, respectively). Item 2 displayed small (and nonsignificant) modification indices, so our choice of this marker item is justified. Freeing the loadings with high modification indices in those countries resulted in acceptable model fit. Although the change in chi-square is still significant ($\Delta\chi^2(66) = 186.4$, $p < .001$), RMSEA and TLI improve compared with the configural invariance model, while the deterioration of CFA is below the 0.01 threshold. Thus, partial metric invariance is satisfied.

Next we tested scalar invariance for those factor loadings that are metrically invariant for the country in question (Steenkamp and Baumgartner 1998). Model fit deteriorated dramatically compared with the partial metric invariance model ($\Delta\chi^2(66) = 1{,}052.9$, $p < .001$; $\Delta\text{RMSEA} = 0.0262$; $\Delta\text{CFI} = -0.068$; $\Delta\text{TLI} = -0.048$). There are numerous large MIs for the item intercepts. We plot the largest MIs in figure 7.

Relaxing these invariance constraints on the item intercepts improves model fit substantially compared with the full scalar invariance model ($\Delta\chi^2(30) = 873.9$, $p < .001$). However, the increase compared with the partial metric invariance model remains significant ($\Delta\chi^2(36) = 179.0$, $p < .001$), and, more important, ΔCFI exceeds the -0.01 threshold. Thus, even partial scalar invariance is not supported. In any case, no items remain that are scalar invariant across all countries. Therefore, within the CFA framework, no means analysis can be undertaken for all countries in the study.

Finally, we tested factor variance invariance, which requires (partial) metric invariance but not scalar invariance (Steenkamp and Baumgartner 1998). Full factor variance invariance is not satisfied ($\Delta\chi^2(10) = 89.3$, $p < .001$). When we relax the constraints on France, Spain, China, Russia, and Japan, we obtain a satisfactory model ($\Delta\chi^2(5) = 9.7$, $p > .05$). Table 6 presents the factor variances.

Comparison of IRT and CFA Results

In this subsection, we compare the substantive insights obtained by our IRT model and the CFA model. Several interesting differences can be observed. In the IRT analysis, it was found that there was substantial variation in the discrimination parameters. However, these results are not mirrored in the CFA factor loadings. Since the discrimination parameters are conceptually similar to factor loadings, the CFA model underestimates the degree of cross-national fluctuation in scale metrics. We conducted a simulation study to further investigate this issue. Detailed results of this study can be obtained from the authors. This finding is consistent with Lubke and Muthén (2004), who find that ignoring the ordinal structure of the data has deleterious effects on tests of measurement invariance.

The estimated country means and variances of SNI based on the IRT analysis are shown in table 6. We also include the SNI variances based on CFA. Note that the latent means cannot be obtained in CFA, because scalar invariance was not supported for any item. Even if scalar invariance would be supported, the general differences in scale usage across countries (see figs. 5 and 6 and table 4) would still make the latent scores problematic to compare.

Consistent with our expectations, the correlation between the IRT-based country mean and a country's score on individualism/collectivism (Hofstede 2001) is significant and negative: $r = -0.734$ ($p < .01$). In addition, there are systematic differences in within-country heterogeneity concerning SNI. Using the IRT results, respondents in relatively

FIGURE 7

LACK OF SCALAR INVARIANCE OF THE SNI SCALE: MODIFICATION INDICES PER COUNTRY

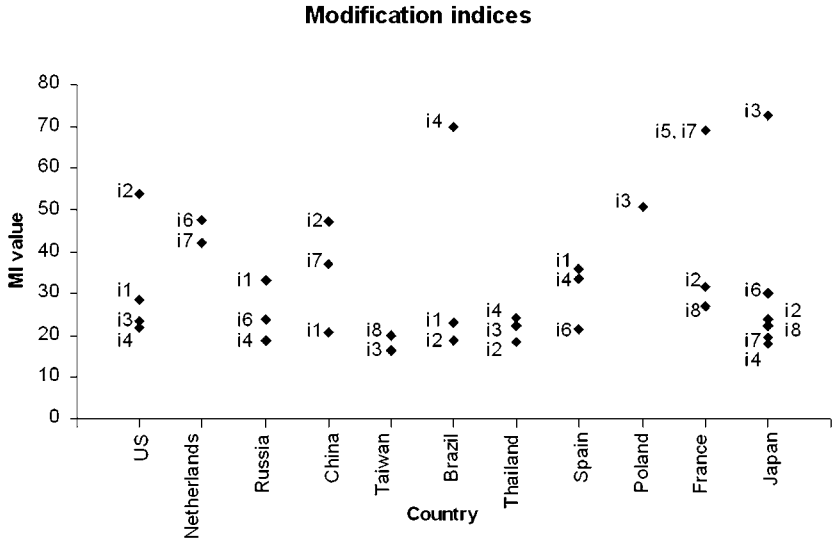


TABLE 6

COUNTRY MEANS AND VARIANCES FOR SNI SCALE

Country	Latent mean IRT	Within-country variance on SNI: IRT model	Within-country variance on SNI: CFA model
France	−1.023	1.106	.458
Netherlands	−1.560	1.408	.325
Spain	−.986	1.462	.463
China	.930	.523	.221
Poland	−.198	.832	.325
Brazil	−.402	.769	.325
Thailand	.136	.730	.325
Russia	.040	.758	.497
United States	−1.200	1.388	.325
Taiwan	.510	.527	.325
Japan	−.216	.809	.415

individualistic countries such as the United States exhibit a higher divergence in their susceptibility to normative influence than respondents in relatively collectivistic countries such as China and Taiwan. As predicted, the correlation between the IRT-based country variance in SNI and national-cultural individualism/collectivism is significant and positive: $r = 0.740$ ($p < .01$). But the correlation between national-cultural individualism/collectivism and the CFA-based latent variable variances is insignificant ($r = 0.02$, $p > .5$), a result that lacks face validity in the light of theory.

IMPLICATIONS FOR CROSS-NATIONAL CONSUMER RESEARCH

We showed that CFA may produce misleading results, that the IRT model appropriately models the ordinal nature of the data, and that measurement invariance restrictions are no longer necessary with the IRT specification. So what are the implications for consumer researchers? First, we advise researchers to be prudent in relying on multigroup CFA:

i) It may produce invalid results due to the ordinal nature of the data;
ii) It runs into problems when the number of groups is large, or more generally, when there are no measurement invariant items;
iii) It can compare subgroups of countries only when invariance is not satisfied, while the IRT model allows a comparison of all countries.

Each of these issues associated with multigroup CFA hinders consumer researchers in deriving substantively meaningful conclusions from cross-national consumer behavior studies.

Many consumer researchers are interested in conducting cross-national mean comparisons and/or comparing structural relationships across countries. How should consumer researchers integrate our approach with their substantive interests? If the goal is to conduct mean comparisons, the IRT model can be used straightaway, since the model provides latent factor scores that are all on the same scale across countries. As we discussed, a Bayesian ANOVA test can be performed to test for mean differences.

If the goal is to compare structural relationships, the optimal way is to model all structural relationships and the IRT measurement part simultaneously. However, such models do not yet exist in either marketing and/or psychometrics because of data limitations and high complexity. As a "second best" option, researchers might use a two-step approach, similar in spirit to Anderson and Gerbing (1988). Such an analysis would require that in a first step, the latent construct scores are estimated. In the second step, these latent scores are used in regression/ANOVA type of techniques to estimate the consumer researcher's substantive hypotheses. Even though simultaneous modeling of measurement and structural model is preferable, the two-step procedure is consistent with the widespread practice in both cross-sectional and experimental research in the social sciences to examine first measurement quality of constructs and thereafter use construct scores (factor scores or summated scores) in regression/ANOVA models (see also Jöreskog [2000], who advocates a similar procedure to deal with interactions and nonlinear effects in LISREL models).

GENERAL DISCUSSION

The dominant logic in consumer behavior research has been that constructs should display certain levels of measurement invariance in order to make valid substantive cross-national comparisons. Indeed, it has been argued that if measurement invariance across countries is lacking, conclusions based on that scale are at best ambiguous and at worst erroneous (Horn 1991). Heeding these recommendations, numerous articles have tested for measurement invariance of constructs, using the multigroup CFA model (e.g., Durvasula et al. 1993; Netemeyer et al. 1991; Steenkamp and Baumgartner 1998; Wong et al. 2003; see Vandenberg and Lance [2000] for an overview of other social sciences). If invariance constraints are not (partially) fulfilled, cross-national comparisons cannot be made. For example, in our application, latent means could not be com-

pared as even partial scalar invariance was not achieved. However, claims of the necessity of certain levels of measurement invariance for particular research objectives are mainly the result of the particular methodology (multigroup CFA) that is used.

In this article we present a model that addresses these problems. Our hierarchical IRT model allows consumer researchers to compare countries substantively despite lack of invariance for any of the items. Moreover, because the ordinal nature of the data is recognized, cross-national differences in scale usage are also accommodated. We found strong noninvariance of scale metrics and of scale usage across countries for SNI. Current CFA-based methodologies are not well suited to account for differences in scale usage because they ignore the ordinal nature of the data (Lubke and Muthén 2004).

Our approach is not limited to studies with many countries. A fixed-effects specification rather than a random-effects specification can be used in studies involving few countries. The estimation procedure would change: the fixed-effects model can be estimated by using noninformative reference priors for the discrimination and threshold parameters in the MCMC procedure. Also, the hierarchical structure for the latent variable can be relaxed in such cases.

There are many issues for further research. Understanding the sources of cross-national differences in discrimination and scale use, as revealed in the discrimination and threshold parameters, is an important topic in its own right. Which cultural processes give rise to cross-national differences in these parameters? The CFA study by Wong et al. (2003) provides an interesting example how studying cross-national differences in response behavior increases our understanding of other cultures. Statistically, covariates that explain variation in item parameters across countries can be incorporated in the IRT measurement model, but more theory and larger data sets are necessary to study relationships between culture and response behavior.

Future work can also focus on extending the current modeling framework in other ways. In consumer research, there is a growing interest in formative measurement models (Jarvis, MacKenzie, and Podsakoff 2003). Future research might extend IRT models—which specify a reflective relation between indicator and construct—to the formative context. In addition, it would be desirable to integrate the IRT measurement model with a hierarchical structural latent variable model that also contains latent predictors. Some recent work has started to address this issue (see Fox 2005b; Fox and Glas 2003), but these models cannot yet accommodate varying item parameters across countries in the measurement models.

Although using the same response format for any specific item across countries is common practice in cross-national consumer research, perhaps it is preferable to use different response formats for any specific item across countries. Like other IRT models (and CFA models), our model allows the scale format to be different for different items but does not allow a different format across countries for the same item. Future research might extend our model to accommodate such differences in response format across countries for similar items, while still arriving at latent scores that are comparable across countries. In equations 3 and 5, the errors could be spatially correlated across countries or correlated via covariates of the countries via a distance measure.

Although many important issues remain for future research, to the best of our knowledge, this is the first article in the social sciences that relaxes all invariance requirements across groups, while retaining the possibility to make substantive comparisons. We hope that our article contributes to stimulating consumer behavior researchers to pay more attention to cross-national measurement issues and thus further advances the rigor of cross-national consumer research.

APPENDIX A
MCMC ALGORITHM

We use Bayesian inference for the IRT model, in which we specify the posterior distribution of all model parameters. For our hierarchical IRT model, the full posterior is given by:

$$
\begin{aligned}
&f(\xi, \gamma, \mathbf{a}, \{\sigma_\gamma^2\}, \{\sigma_a^2\}, \{\sigma_g^2\}, \tau^2 \mid x)\\
&\propto \prod_g \left[\prod_i \left[\prod_k [\, f(x_{ik}^g \mid \xi_i^g, \gamma_k^g, a_k^g)] f(\xi_i^g \mid \xi^g, \sigma_g^2) \right] f(\xi^g \mid \tau^2) \right]\\
&\quad \times \prod_g \left[\prod_k [\, f(\gamma_k^g \mid \gamma_k, \sigma_{\gamma_k}^2)] f(a_k^g \mid a_k, \sigma_{a_k}^2) \right]\\
&\quad \times \prod_g [f(\sigma_g^2)] \prod_k [\, f(\sigma_{\gamma_k}^2) f(\sigma_{a_k}^2) f(\gamma_k) f(a_k)] f(\tau^2).
\end{aligned}
$$

We use data augmentation (Tanner and Wong 1987) to facilitate estimation. A Metropolis-Hastings step is used to sample the threshold parameters, for which the full conditional distribution is complex. The Gibbs sampler consists of the following steps:

1. Sample from $[Z_{ik}^g \mid X_{ik}^g, \xi_i^g, a_k^g, \gamma_k^g]$, for $k = 1, \ldots, K$, $i = 1, \ldots, N_g$, and $g = 1, \ldots, G$.

   Given the variables X_{ik}^g, ξ_i^g, a_k^g, and γ_k^g, the variables Z_{ik}^g are independent and normally distributed: $Z_{ik}^g \mid X_{ik}^g, \xi_i^g, a_k^g, \gamma_k^g \sim N(a_k^g \xi_i^g, 1)\, I(\gamma_{k,c-1}^g < Z_{ik}^g < \gamma_{k,c}^g)$ if $X_{ik}^g = c$.
2. Sample from $[\xi_i^g \mid Z_{ik}^g, a_k^g, \xi^g, \sigma_g^2]$, for $i = 1, \ldots, N_g$, and $g = 1, \ldots, G$.

   The full conditional distribution is a product of two normal distributions, and from standard properties it follows that:

$$
\xi_i^g \mid \boldsymbol{Z}_i^g, \boldsymbol{a}^g, \xi^g, \sigma_g^2
$$

$$
\sim N\left(\frac{\sum_{k=1}^K a_k^g Z_{ik}^g + \xi^g/\sigma_g^2}{\sum_{k=1}^K (a_k^g)^2 + \sigma_g^{-2}},\ \frac{1}{\sigma_g^{-2} + \sum_{k=1}^K (a_k^g)^2} \right).
$$

3. Sample from $[a_k^g \mid \xi^g, \boldsymbol{Z}_k^g, a_k, \sigma_{a_k}^2]$, $g = 1, \ldots, G$, $k = 1, \ldots, K$.

The prior is $a_k^g \sim N(a_k, \sigma_{a_k}^2) I(a_k^g \in (0, A])$. Therefore, the full posterior is normal, with

$$a_k^g \mid \xi^g, \boldsymbol{Z}_k^g, a_k, \sigma_{a_k}^2$$

$$\sim N\left(\frac{\sum_{i=1}^{N_g} \xi_i^g Z_{ik}^g + a_k/\sigma_{a_k}^2}{\sum_{i=1}^{N_g} (\xi_i^g)^2 + \sigma_{a_k}^{-2}}, \frac{1}{\sum_{i=1}^{N_g} (\xi_i^g)^2 + \sigma_{a_k}^{-2}}\right) I(a_k^g \in (0, A]),$$

where A is a positive number. For identification, it is imposed that $\prod_{k=1}^{K} a_k^g = 1$.

4. Sample from $[\gamma_k^g \mid \gamma_k, \sigma_{\gamma_k}^2, a_k^g, Z_{ik}^g, X_{ik}^g]$, $g = 1, \ldots, G$, $k = 1, \ldots, K$, and $c = 1, \ldots, C - 1$.

The full conditional posterior of the threshold parameters is proportional to:

$$\prod_{i \mid g} P(\gamma_{k,x_{ik}^g}^g > Z_{ik}^g > \gamma_{k,x_{ik}^g-1}^g \mid \xi_i^g, a_k^g, \gamma_k^g) f(\gamma_k^g \mid \gamma_k, \sigma_{\gamma_k}^2).$$

A Metropolis-Hastings algorithm is used to simulate a realization from this posterior distribution. In the mth iteration of the MCMC chain we draw a candidate $\gamma_k^{g,*}$ from (Fox 2005b):

$$\gamma_{k,c}^{g,*} \sim N(\gamma_{k,c}^{g,m-1}, \sigma_{MH}^2) I(\gamma_{k,c-1}^{g,*} < \gamma_{k,c}^{g,*} < \gamma_{k,c+1}^{g,m-1})$$

for $c = 1, \ldots, C - 1$, where σ_{MH}^2 is a tuning parameter to adjust the accept/reject rate of the algorithm. The Metropolis-Hastings acceptance probability is then given by:

$$\min\left[\prod_{i \mid g} \frac{\Pr(X_{ik}^g = x_{ik}^g \mid \xi_i^g, a_k^g, \gamma_k^{g,*})}{\Pr(X_{ik}^g = x_{ik}^g \mid \xi_i^g, a_k^g, \gamma_k^{g,m-1})} \frac{f(\gamma_k^{g,*} \mid \gamma_k, \sigma_{\gamma_k}^2)}{f(\gamma_k^{g,m-1} \mid \gamma_k, \sigma_{\gamma_k}^2)} \frac{f(\gamma_k^{g,m-1} \mid \gamma_k^{g,*}, \sigma_{MH}^2)}{f(\gamma_k^{g,*} \mid \gamma_k^{g,m-1}, \sigma_{MH}^2)}, 1\right].$$

The first two parts of this expression represent the contribution from the likelihood, the second part comes from the proposal distributions. For identification, we set $\sum_{k=1}^{K} \gamma_{k3}^g = 0$.

5. Sample from $[a_k \mid a_k^g, \sigma_{a_k}^2]$, and $[\gamma_k \mid \gamma_k^g, \sigma_{\gamma_k}^2]$, for $k = 1, \ldots, K$.

We use a Metropolis-Hastings algorithm to sample these parameters.

6. Sample from $[\sigma_{a_k}^2 \mid a_k^g, a_k]$ for $k = 1, \ldots, K$, and $[\sigma_g^2 \mid \xi_i^g, \xi^g]$, $[\tau^2 \mid \xi^g]$.

For each variance parameter an inverse gamma prior is specified with parameters g_1 and g_2. As a result, each full conditional (except for $\sigma_{\gamma_k}^2$) has an inverse gamma distribution with shape parameter $G/2 + g_1$ for each of the K items, $N_g/2 + g_1$, and $G/2 + g_1$, respectively, and scale parameter

$$g_2 + \sum_{g=1}^{G} (a_k^g - a_k)^2/2, \; g_2 + \sum_{i=1}^{N_g} (\xi_i^g - \xi^g)^2/2,$$

$$g_2 + \sum_{g=1}^{G} (\xi^g)^2/2,$$

respectively. Noninformative proper priors were specified with $g_1 = g_2 = 1$. For $\sigma_{\gamma_k}^2$, a Metropolis-Hastings algorithm was used with an inverse gamma prior with parameters 1 and 1.

7. Sample from $[\xi^g \mid \sigma_g^2, \tau^2]$ for $g = 1, \ldots, G$. The prior is $\xi^g \sim N(\xi, \tau^2)$, so that

$$\xi^g \mid \xi, \tau^2, \sigma_g^2 \sim N\left(\frac{\sum_{i \mid g} \xi_i^g/\sigma_g^2 + \xi/\tau^2}{N_g/\sigma_g^2 + \tau^{-2}}, \frac{1}{N_g/\sigma_g^2 + \tau^{-2}}\right).$$

8. Sample from $[\xi \mid \xi^g, \tau^2]$. With a noninformative prior, we have

$$\xi \mid \xi^g, \tau^2 \sim N\left(G^{-1} \sum_g \xi^g, G^{-1}\tau^2\right).$$

APPENDIX B

BAYESIAN TESTS

In order to test factor variance invariance, consider $G - 1$ linearly independent contrasts $\Delta_g = \log \sigma_g^2 - \log \sigma_G^2$. Then, the hypothesis $\Delta_0 = 0$ corresponds with equal factor variances across groups. The density function $p(\Delta \mid \mathbf{x})$ is a monotonic decreasing function of a function Q_0, which is asymptotically distributed as χ_{G-1}^2, as $N_g \to \infty$ (see Box and Tiao 1973). Hence, for large samples, the vector $\Delta_0 = 0$ is included in the highest posterior density (HPD) region of $1 - \alpha$ if and only if:

$$\lim_{N_g \to \infty} P[p(\Delta \mid \mathbf{x}) > p(\Delta_0 \mid \mathbf{x}) \mid \mathbf{x}]$$

$$= P\left(\chi_{G-1}^2 < \frac{Q_0}{1 + A}\right) < 1 - \alpha, \qquad \textbf{(B1)}$$

where

$$Q_0 = -\sum_{g=1}^{G} N_g(\log s_g^2 - \log \bar{s}^2)$$

$$A = \frac{1}{3(G-1)}\left(\sum_{g=1}^{G} N_g^{-1} - N^{-1}\right)$$

and s_g^2 and $\bar{s}^2$ are the mean sum of squares in group G and the overall mean sum of squares, respectively. N_g is the sample size in country g, and $N = \Sigma N_g$. The MCMC algorithm can be used to compute the right-hand side of equation B1 (see Fox 2005a). It follows that the hypothesis of equal variances across countries is rejected when

$$P\left(\chi^2_{G-1} < \frac{Q_0}{1+A}\right) > 1 - \alpha.$$

For latent means, we can consider $G - 1$ linear contrasts $\Delta_g = \xi^g - \xi^G$. Then, it holds that $p(\Delta \mid \mathbf{x})$ is a monotonic decreasing function of a function Q_0, which is asymptotically distributed as $F_{(G-1,N-G)}$ as $N_g \to \infty$. For large samples, the vector $\Delta_0 = 0$ is included in the highest posterior density region of $1 - \alpha$ if and only if:

$$\lim_{N_g \to \infty} P[p(\Delta \mid \mathbf{x}) > p(\Delta_0 \mid \mathbf{x}) \mid \mathbf{x}]$$

$$= P\left(F_{(G-1,N-G)} < \frac{\sum_g N_g(\xi^g - \bar{\xi})^2}{(G-1)s^2}\right) < 1 - \alpha,$$

where $\bar{\xi} = \frac{1}{N}\sum_g N_g \xi^g$. Again the hypothesis of equal means across countries is rejected when

$$P\left(F_{(G-1,N-G)} < \frac{\sum_g N_g(\xi^g - \bar{\xi})^2}{(G-1)s^2}\right) > 1 - \alpha.$$

REFERENCES

Alden, Dana L., Jan-Benedict E. M. Steenkamp, and Rajeev Batra (2006), "Consumer Attitudes toward Marketplace Globalization: Structure, Antecedents, and Consequences," *International Journal of Research in Marketing*, 23 (September), 227–39.

Anderson, James C. and David W. Gerbing (1988), "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach," *Psychological Bulletin*, 103 (May), 411–23.

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.

Balasubramanian, Siva K. and Wagner A. Kamakura (1989), "Measuring Consumer Attitudes toward the Marketplace with Tailored Interviews," *Journal of Marketing Research*, 26 (August), 311–26.

Batra, Rajeev, Venkatram Ramaswamy, Dana L. Alden, Jan-Benedict E. M. Steenkamp, and S. Ramachander (2000), "Effects of Brand Local and Nonlocal Origin on Consumer Attitudes in Developing Countries," *Journal of Consumer Psychology*, 9, 83–95.

Baumgartner, Hans (2004), "Issues in Assessing Measurement Invariance in Cross-National Research," presentation at Sheth Foundation/Sudman Symposium on Cross-Cultural Survey Research, University of Illinois, Urbana-Champaign.

Baumgartner, Hans and Jan-Benedict E. M. Steenkamp (1998), "Multi-Group Latent Variable Models for Varying Numbers of Items and Factors with Cross-National and Longitudinal Applications," *Marketing Letters*, 9 (1), 21–35.

——— (2001), "Response Styles in Marketing Research: A Cross-National Investigation," *Journal of Marketing Research*, 38 (May), 143–56.

Bearden, William O., Richard G. Netemeyer, and Jesse E. Teel (1989), "Measurement of Consumer Susceptibility to Interpersonal Influence," *Journal of Consumer Research*, 15 (March), 473–81.

——— (1990), "Further Validation of the Consumer Susceptibility to Interpersonal Influence Scale," in *Advances in Consumer Research*, Vol. 17, ed. Marvin E. Goldberg, Gerald Gorn, and Richard W. Pollay, Provo, UT: Association for Consumer Research, 770–76.

Bechtel, Gordon G. (1985), "Generalizing the Rasch Model for Consumer Rating Scales," *Marketing Science*, 4 (Winter), 62–73.

Berger, James O. and Mohan Delampady (1987), "Testing Precise Hypotheses," *Statistical Science*, 2, 317–52.

Best, Nicky, Kate Cowles, and Karen Vines (1995), *CODA Convergence Diagnosis and Output Analysis Software for Gibbs Sampler Output: Version 0.3*, Computer software and manual, Cambridge: MRC Biostatistics Unit, Institute of Public Health.

Bolt, Daniel M., Jennifer E. Hare, Jennifer E. Vitale, and Joseph P. Newman (2004), "A Multigroup Item Response Theory Analysis of the Psychopathy Checklist—Revised," *Psychological Assessment*, 16 (2), 155–68.

Box, George E. P. and George C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Bradlow, Eric T. and Gavan J. Fitzsimons (2001), "Subscale Distance and Item Clustering Effects in Self-Administered Surveys: A New Metric," *Journal of Marketing Research*, 38 (May), 254–61.

Burisch, Matthias (1984), "Approaches to Personality Inventory Construction," *American Psychologist*, 39 (March), 214–27.

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 (3), 456–66.

Cheung, Gordon W. and Roger B. Rensvold (2002), "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance," *Structural Equation Modeling*, 9 (2), 233–55.

De Jong, Martijn G., Jan-Benedict E. M. Steenkamp, Jean-Paul Fox, and Hans Baumgartner (2007), "Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation," *Journal of Marketing Research*, forthcoming.

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), 626–36.

Fox, Jean-Paul (2005a), "Multilevel IRT Model Assessment," in *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*, ed. Andries van der Ark, Marcel A. Croon, and Klaas Sijtsma, London: Erlbaum, 227–52.

——— (2005b), "Multilevel IRT Using Dichotomous and Polytomous Items," *British Journal of Mathematical and Statistical Psychology*, 58, 145–72.

Fox, Jean-Paul and Cees A. W. Glas (2001), "Bayesian Estimation of a Multilevel IRT Model Using Gibbs Sampling," *Psychometrika*, 66 (June), 269–86.

——— (2003), "Bayesian Modeling of Measurement Error in Predictor Variables Using Item Response Theory," *Psychometrika*, 68 (June), 169–92.

Franses, Philip Hans and Richard Paap (2001), *Quantitative Models in Marketing Research*, Cambridge: Cambridge University Press.

Greene, William H. (2003), *Econometric Analysis*, Upper Saddle River, NJ: Prentice-Hall.

Hofstede, Geert H. (2001), *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Nations*, 2nd ed., Thousand Oaks, CA: Sage.

Holland, Paul W. and Howard Wainer (1993), *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

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, 114–25.

Horn, John L. and J. Jack McArdle (1992), "A Practical and Theoretical Guide to Measurement Invariance in Aging Research," *Experimental Aging Research*, 18 (Fall–Winter), 117–44.

Janssen, Rianne, Francis Tuerlinckx, Michel Meulders, and Paul de Boeck (2000), "A Hierarchical IRT Model for Criterion-Referenced Measurement," *Journal of Educational and Behavioral Statistics*, 25 (Fall), 285–306.

Jarvis, Cheryl, Scott B. MacKenzie, and Philip M. Podsakoff (2003), "A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research," *Journal of Consumer Research*, 30 (September), 199–218.

Jöreskog, Karl G. (2000), "Latent Variable Scores and Their Uses," working paper, Scientific Software International, Chicago.

Kagitcibasi, Cigdem (1997), "Individualism and Collectivism," in *Social Behavior and Applications*, Vol. 3 of *Handbook of Cross-Cultural Psychology*, 2nd ed., ed. John W. Berry, Marshall H. Segall, and Cigdem Kagitcibasi, Boston: Allyn & Bacon, 1–49.

Kass, Robert E. and Adrian E. Raftery (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90 (June), 773–95.

Lord, Frederic M. and Melvin R. Novick (1968), *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

Lubke, Gitta H. and Bengt O. Muthén (2004), "Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons," *Structural Equation Modeling*, 11 (4), 514–34.

MacCallum, Robert C., Mary Roznowski, and Lawrence B. Necowitz (1992), "Model Modifications in Covariance Structure Analysis: The Problem of Capitalization on Chance," *Psychological Bulletin*, 111 (3), 490–504.

MacKenzie, Scott B. (2003), "The Dangers of Poor Construct Conceptualization," *Journal of the Academy of Marketing Science*, 31 (Summer), 323–26.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

Mangleburg, Tamara F. and Terry Bristol (1998), "Socialization and Adolescents' Skepticism toward Advertising," *Journal of Advertising*, 27 (Fall), 11–21.

May, Henry (2006), "A Multilevel Bayesian IRT Method for Scaling Socioeconomic Status in International Studies of Education," *Journal of Educational and Behavioral Statistics*, 31 (Spring), 63–79.

McCracken, Grant (1986), "Culture and Consumption: A Theoretical Account of the Structure and Movement of the Cultural Meaning of Consumer Goods," *Journal of Consumer Research*, 13 (June), 71–84.

Meade, Adam W. and Gary J. Lautenschlager (2004), "A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance," *Organizational Research Methods*, 7 (October), 361–88.

Meredith, William (1993), "Measurement Invariance, Factor Analysis, and Factorial Invariance," *Psychometrika*, 58 (December), 525–43.

——— (1995), "Two Wrongs May Not Make a Right," *Multivariate Behavioral Research*, 30 (1), 89–94.

Mick, David Glen (1996), "Are Studies of Dark Side Variables Confounded by Socially Desirable Responding? The Case of Materialism," *Journal of Consumer Research*, 23 (September), 106–19.

——— (2005), "Meaning and Mattering through Transformative Consumer Research," Presidential Address at the Conference of the Association for Consumer Research, San Antonio, September 28–30.

Monroe, Kent B. (1993), "Editorial," *Journal of Consumer Research*, 19 (March), v.

Netemeyer, Richard G., Srinivas Durvasula, and Donald R. Lichtenstein (1991), "A Cross-National Assessment of the Reliability and Validity of the CETSCALE," *Journal of Marketing Research*, 28 (August), 320–27.

Newton, Michael E. and Adrian E. Raftery (1994), "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society B*, 56 (1), 3–48.

Novak, Thomas P., Donna L. Hoffman, and Yiu-Fai Yung (2000), "Measuring the Customer Experience in Online Environments: A Structural Modeling Approach," *Marketing Science*, 19 (Winter), 22–42.

Raju, Nambury S., Barbara M. Byrne, and Larry J. Laffitte (2002), "Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory," *Journal of Applied Psychology*, 87 (3), 517–29.

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 (3), 552–66.

Richins, Marsha L. (1994), "Special Possessions and the Expression of Material Values," *Journal of Consumer Research*, 21 (December), 522–33.

Richins, Marsha L. and Scott Dawson (1992), "A Consumer Values Orientation for Materialism and Its Measurement: Measure

Development and Validation," *Journal of Consumer Research*, 19 (December), 303–16.

Rossi, Peter E., Zvi Gilula, and Greg M. Allenby (2001), "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach," *Journal of the American Statistical Association*, 96 (March), 20–31.

Roth, Martin S. (1995), "The Effects of Culture and Socioeconomics on the Performance of Global Brand Image Strategies," *Journal of Marketing Research*, 32 (May), 163–75.

Samejima, Fumiko (1969), "Estimation of Latent Ability Using a Response Pattern of Graded Scores," *Psychometrika Monograph Supplement*, 17, 1–100.

——— (1972), "A General Model for Free Response Data," *Psychometrika Monograph Supplement*, 18, 1–68.

Schwartz, Shalom H. (1994), "Cultural Dimensions of Values: Toward an Understanding of National Differences," in *Individualism and Collectivisim: Theory, Method and Application*, ed. Uichol Kim, Harry C. Triandis, Cigdem Kagitcibasi, Sang-Chin Choi, and Gene Yoon, Thousand Oaks, CA: Sage, 85–119.

Sen, Sankar, Zeynep Gürhan-Canli, and Vicki G. Morwitz (2001), "Withholding Consumption: A Social Dilemma Perspective on Consumer Boycotts," *Journal of Consumer Research*, 28 (December), 399–417.

Singh, Jagdip, Roy D. Howell, and Gary K. Rhoads (1990), "Adaptive Designs for Likert-Type Data: An Approach for Implementing Marketing Surveys," *Journal of Marketing Research*, 27 (August), 304–21.

Steenkamp, Jan-Benedict E. M. (2005), "Moving Out of the U.S. Silo: A Call to Arms for Conducting International Marketing Research," *Journal of Marketing*, 69 (October), 6–8.

Steenkamp, Jan-Benedict E. M. and Hans Baumgartner (1998), "Assessing Measurement Invariance in Cross-National Consumer Research," *Journal of Consumer Research*, 25 (June), 78–90.

Steenkamp, Jan-Benedict E. M. and Katrijn Gielens (2003), "Consumer and Market Drivers of the Trial Rate of New Consumer Products," *Journal of Consumer Research*, 30 (December), 368–84.

Tanner, Martin A. and Wing Hung Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82 (June), 528–50.

Thissen, David, Lynne Steinberg, and Meg Gerrard (1986), "Beyond Group Mean Differences: The Concept of Item Bias," *Psychological Bulletin*, 99 (1), 118–28.

Thissen, David, Lynne Steinberg, and Howard Wainer (1988), "Use of Item Response Theory in the Study of Group Differences in Trace Lines," in *Test Validity*, ed. Howard Wainer and Henry I. Braun, Hillsdale, NJ: Erlbaum, 147–69.

——— (1993), "Detection of Differential Item Functioning Using the Parameters of Item Response Models," in Holland and Wainer 1993, 67–113.

Triandis, Harry C. (1989), "The Self and Social Behavior in Differing Cultural Contexts," *Psychological Review*, 96 (July), 506–20.

Tse, David K., Kam-Hon Lee, Ilan Vertinsky, and Donald A. Wehrung (1988), "Does Culture Matter? A Cross-Cultural Study of Executives' Choice, Decisiveness, and Risk Adjustment in International Marketing," *Journal of Marketing*, 52 (October), 81–95.

Vandenberg, Robert J. and Charles E. Lance (2000), "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research," *Organizational Research Methods*, 3 (1), 4–70.

Weathers, Danny, Subhash Sharma, and Ronald W. Niedrich (2005), "The Impact of the Number of Scale Points, Dispositional Factors, and the Status Quo Decision Heuristic on Scale Reliability and Response Accuracy," *Journal of Business Research*, 58 (11), 1516–24.

Welkenhuysen-Gybels, Jerry, Jaak Billiet, and Bart Cambré (2003), "Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items," *Journal of Cross-Cultural Psychology*, 34 (November), 702–22.

Wong, Nancy, Aric Rindfleisch, and James E. Burroughs (2003), "Do Reverse-Worded Items Confound Measures in Cross-Cultural Consumer Research? The Case of the Material Values Scale," *Journal of Consumer Research*, 30 (June), 72–91.

Wooten, David B. and Americus Reed (2004), "Playing It Safe: Susceptibility to Normative Influence and Protective Self-Presentation," *Journal of Consumer Research*, 31 (December), 551–56.