MARKETING SCIENCE
Vol. 28, No. 4, July–August 2009, pp. 674–689
ISSN 0732-2399 | EISSN 1526-548X | 09 | 2804 | 0674

informs

DOI 10.1287/mksc.1080.0439

A Model for the Construction of Country-Specific Yet Internationally Comparable Short-Form Marketing Scales

Martijn G. de Jong
Department of Marketing Management, Rotterdam School of Management, Erasmus University, 3062 PA Rotterdam, The Netherlands, mjong@rsm.nl

Jan-Benedict E. M. Steenkamp
Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, jbs@unc.edu

Bernard P. Veldkamp
Department of Research Methodology, Measurement and Data Analysis, University of Twente, 7500 AE Enschede, The Netherlands, b.p.veldkamp@gw.utwente.nl

In the last few decades, the measurement of marketing constructs has improved tremendously. Our discipline has also started to systematically catalogue our measurement knowledge in influential handbooks of marketing scales. However, at least two important issues remain. First, existing scales are often too long for administration in nonstudent samples or in applied studies. Second, existing (U.S.-developed) scales may contain items that are not informative about the underlying construct in particular countries, whereas relevant items tapping into local cultural expressions of the construct in question may be missing. To address these issues, we propose a new model that yields country-specific yet fully cross-nationally comparable short forms of unidimensional marketing scales. The procedure is based on hierarchical item response theory and optimal test design. The procedure is flexible in the sense that the researcher can specify various constraints on item content, scale length, and measurement precision. Because our procedure allows inclusion of country-specific (or "emic") items in standardized (or "etic") scales, it presents an important step toward resolving the emic-etic dilemma that has plagued international marketing research for decades.

Key words: measurement; marketing research; marketing surveys; international marketing; scale construction; measurement invariance
History: Received: January 18, 2007; accepted: March 21, 2008; processed by Eric T. Bradlow. Published online in *Articles in Advance* February 12, 2009.

1. Introduction

Using scientific methods of measurement, analysis, and interpretation is the foundation of marketing's claim to be a science. Consequently, it is encouraging that marketing scientists have made tremendous progress in accurately measuring marketing constructs. Much of this knowledge has been systematically collected in handbooks of marketing scales such as Bearden and Netemeyer (1999) and the series published by the American Marketing Association (Bruner and Hensel 1992, and successive volumes). The popularity of these books reflects marketing science's strong desire to use rigorously validated measurement instruments rather than ad hoc-constructed scales. Despite this significant progress, several issues remain unresolved.

First, many scales are too long to be useful for effective administration, especially in nonstudent samples (Bergkvist and Rossiter 2007). Long scales lead to high costs of data collection and respondent fatigue, frustration, and attrition (Benet-Martínez and John 1998). Consequently, researchers have started to develop short forms of existing scales (e.g., Richins 2004, Shimp and Sharma 1987, Steenkamp and Baumgartner 1995). However, because the classical test theory techniques used to select items yield sample-dependent item characteristics, one cannot simply use item characteristics obtained in the calibration sample (e.g., factor score coefficients) to construct latent scores on future samples. Moreover, existing common practice to select high-loading items for the short form does not allow the researcher to measure particular ranges of the latent construct with more precision (e.g., high satisfaction or loyalty), even if academic insight dictates otherwise (Fraley et al. 2000, Gupta and Zeithaml 2006).

Second, most marketing scales have been developed and tested only in the United States. Various scholars have urged marketing scientists to conduct more research on an international basis (Bolton 2003, Shugan 2006, Steenkamp 2005, Winer 1998). Unfortunately, scale length becomes an even more pertinent issue as data collection costs multiply by the number of countries. The educational attainment of non-U.S. respondents is often lower, making respondent fatigue and attrition even more problematic. The rigorous, unidimensional properties of U.S. scales may not be fully upheld in other countries. Furthermore, U.S.-developed scale items may not be equally informative in other countries (Thompson 2007). Ever since Berry's (1969) seminal article, international researchers in the social sciences have argued that country-specific ("emic") items might have to be added to or replace cross-nationally standardized ("etic") items (Aaker et al. 2001, Kumar 2000, Steenkamp 2005). Thus, valid measurement may require (some) country-specific items, whereas cross-nationally comparable measurement requires standardized scales. This emic-etic dilemma has been called "one of the major problems faced by an international marketing researcher" (Kumar 2000, p. 129).

In this paper, we propose an integrated methodology that addresses these issues. Our procedure yields country-specific short-form marketing scales, subject to researcher-specified optimization constraints such as varying measurement precision across the latent continuum. It controls for deviations from unidimensionality in specific countries by allowing for excess correlations between items. It provides a measure of fit per country and yields short-form latent scores that are comparable *within* countries across samples as well as *between* countries. Our procedure is based on a combination of two powerful psychometric tools: hierarchical item response theory (De Boeck and Wilson 2004, Johnson et al. 2006) and optimal test design methods (Van der Linden 2005).

2. A Model for the Construction of Short-Form Marketing Scales

2.1. Item Response Theory vs. Classical Test Theory for Scale Construction

Construction and evaluation of marketing scales has relied heavily on classical test theory (CTT) and its most important statistic, Cronbach's alpha, which is the lower bound on the scale's reliability (Netemeyer et al. 2003). In contrast, our model is based on item response theory (IRT) (Lord and Novick 1968). IRT has several important advantages over CTT, which will be important for our purposes. First, in CTT, measurement error is assumed to be constant across the entire range of the latent trait. In IRT, measurement error is allowed to vary across levels of the underlying construct. Second, in CTT, reliability is a joint property of all items in the scale and the particular individuals sampled. When items are added to or dropped from a scale in CTT, the usefulness of the other items to the quality of the scale will change. In IRT, items are posited to contribute independently to measurement precision. When measurement precision is not sufficiently accurate at certain levels of the construct, items can be added that increase the precision at those levels.

Third, in CTT, reliability estimates vary across samples because it is a function of sample homogeneity. In IRT, measurement precision of items is theoretically invariant across samples. Finally, to allow comparisons among respondents, CTT dictates that all respondents answer all items.[1] IRT has item-free calibration, which implies that respondents who have answered different questions can still be compared provided that the items have all been calibrated on a common scale and are stored in an *item bank* that contains the item parameters describing the items. The unique, item-free calibration feature of IRT will be of paramount importance for constructing country-specific scales while retaining cross-national comparability.

2.2. A Graded Response IRT Model for Cross-National Research

Our point of departure is a unidimensional set of Likert items, arguably the standard scale format for marketing scales (Bearden and Netemeyer 1999). The most suitable IRT model for such data is Samejima's (1969) graded response model. de Jong et al. (2007) extended the Samejima (1969) model to a multicountry setting. Their model can be written as

$$\Pr(X_{ik}^g = c) = \Phi(a_k^g \xi_i^g - \gamma_{k,c-1}^g) - \Phi(a_k^g \xi_i^g - \gamma_{k,c}^g), \quad (1)$$

$$\gamma_{k,c}^g \sim N(\gamma_{k,c}, \sigma_\gamma^2) \quad \text{for } c = 1, \ldots, C-1,$$

$$\gamma_{k,1}^g \leq \cdots \leq \gamma_{k,C-1}^g,\ \gamma_{k,1} \leq \cdots \leq \gamma_{k,C-1}, \quad (2)$$

$$\log a_k^g \sim N(\log a_k, \sigma_a^2), \quad (3)$$

$$\xi_i^g \sim N(\xi^g, \sigma_g^2), \quad (4)$$

$$\xi^g \sim N(\xi, \tau^2), \quad (5)$$

where country is indexed by g, respondent is indexed by i ($i = 1, \ldots, N_g$), item is indexed by k ($k = 1, \ldots, K_g$), response option is indexed by c ($c = 1, \ldots, C$),

[1] In theory, different respondents could rate different subsets of items. Each element of the covariance matrix is based on those respondents with complete data for the pair of items. However, CTT researchers caution against this as it is prone to serious psychometric problems (e.g., nonpositive definite matrices and standardized covariances exceeding unity) (Bollen 1989).

and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Equation (1) represents the probability of choosing a particular response category as a function of the underlying latent trait and the item parameters. Equation (2) 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. Analogously, Equation (3) posits that the discrimination parameter a_k^g is the sum of an overall mean discrimination parameter and country-specific deviation. The discrimination parameter should be positive. Hence, negatively worded items should be reverse coded before applying the model. The latent trait 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.

2.3. Incorporating Emic Items

The model proposed by de Jong et al. (2007) is only applicable to standardized ("etic") scales. However, valid measurement in different countries may require that existing scales are modified to different country-cultural contexts (Aaker et al. 2001, Burgess and Steenkamp 2006). The question is how to compare country scores if a scale contains both etic and emic items.

One way to address this issue is to use the IRT model proposed by May (2006) in a cross-national context. May's (2006) model allows researchers to collect different items in different countries as long as there are also common items. However, his model requires that measurement invariance be imposed for the common items. In settings with many countries, invariance is unlikely to be satisfied for *any* item (Baumgartner 2004).

We propose a different solution by extending the structure displayed in (2) and (3). Our solution allows for country-specific items *without requiring any item to be invariant*. The formulation becomes

$$\gamma_{k,c}^g \sim N(\gamma_{k,c}, \sigma_\gamma^2) \quad \text{for } c=1,\ldots,C-1, \gamma_{k,1}^g \le \cdots \le \gamma_{k,C-1}^g, \quad \text{if } k \in \Xi, \tag{6}$$

$$\log a_k^g \sim N(\log a_k, \sigma_a^2) \quad \text{if } k \in \Xi, \tag{7}$$

$$\gamma_{k,c}^g \text{ uniform}, \quad \text{subject to} \quad \gamma_{k,1}^g \le \cdots \le \gamma_{k,C-1}^g \quad \text{if } k \in \Theta_g, \tag{8}$$

$$\log a_k^g \sim N(\mu_a, V_a) \quad \text{if } k \in \Theta_g, \tag{9}$$

where Ξ is the set of etic items and Θ_g is the set of emic items unique to country g. For country-specific items, we impose a vague lognormal prior on the discrimination parameter (i.e., μ_a and V_a are chosen so that the prior is vague). A uniform prior is chosen for the threshold parameters subject to inequality constraints (Johnson and Albert 1999). The number of items can vary across countries. It is necessary to have etic items to calibrate the model, but these common items do *not* have to be invariant across countries.

2.4. Accounting for Deviations from Unidimensionality

Thus far, we have assumed that the set of items is unidimensional. Unidimensionality refers to the requirement that one latent construct can account for the covariation between the items (Gerbing and Anderson 1988). However, excess covariation between items is not uncommon even for established scales because of, for example, scale length or item wording (Baumgartner and Steenkamp 2006). Cross-national research adds additional sources of excess covariation such as cultural variations and issues in translation (Kumar 2000, Thompson 2007). We allow for subsets of items to exhibit residual dependencies, conditional on the latent substantive trait, by including person-specific "testlet" effects $\psi_{i,d_k^g}^g$ that deal with common stimulus elements in items (Bradlow et al. 1999; see also Tuerlinckx and De Boeck 2004):

$$\Pr(X_{ik}^g = c) = \Phi\left(a_k^g\left(\xi_i^g - \psi_{i,d_k^g}^g\right) - \gamma_{k,c-1}^g\right) - \Phi\left(a_k^g\left(\xi_i^g - \psi_{i,d_k^g}^g\right) - \gamma_{k,c}^g\right), \tag{10}$$

$$\psi_{i,d}^g \sim N\left(0, \sigma_{\psi_d^g}^2\right), \tag{11}$$

where d_k^g indicates the subset of item k in country g (it is assumed that there are D^g subsets in country g). The subset selection parameter is indexed by country, which implies that the subsets of items exhibiting excess covariation need not be the same across countries. The variance of the testlet effects is allowed to vary across subsets and countries. Normally, most items have zero excess correlations, $\psi_{i,d_k^g}^g = 0$, in that they are independent conditional on the latent trait. The additional parameters are formulated as a deviation from a person's trait value so that there is a systematic increase or decrease in the trait value for the items in the same subset: $\psi_{i,d}^g \sim N(0, \sigma_{\psi_d^g}^2)$. The parameter affects all items in the subset and is constant across the subset, so that the residual dependencies among the items are incorporated (Bradlow et al. 1999, Wang et al. 2002).

2.5. Estimation and Fit

Identification requires that mean and variance of the latent scale be fixed in every country. Both a uniform shift in threshold parameters in Equations (6) and (8) as well as a shift in the latent mean in Equation (4) can capture the country mean of the latent scale. Similarly, for the variance, we can uniformly shift the discrimination parameters in Equations (7) and (9) as well

as the latent variance in Equation (4). To identify the model, we impose

$$\prod_{k\in\Xi,\Theta_g} a_k^g = 1, \qquad \sum_{k\in\Xi,\Theta_g} \gamma_{k,3}^g = 0. \tag{12}$$

The model is estimated via Markov chain Monte Carlo (MCMC) techniques. Estimation details are given in the appendix. Model fit has traditionally received relatively little attention in IRT. However, in a cross-national context in which the scale is administered in various languages to different cultures, the extent to which the IRT model fits the data in each country becomes especially pertinent. We will consider *overall* fit as well as the fit of *specific* aspects of the model.

2.5.1. Overall Fit. If an IRT model fits the data, the distribution of posterior scores should overlap with the distribution of the observed sum scores (Béguin and Glas 2001). The observed scores of all persons and items are stored in the matrix $\mathbf{X}$ and all model parameters are stored in the vector $\boldsymbol{\omega}$. In each MCMC iteration after the burn-in phase, a new replicated data set, $\mathbf{X}_{\text{rep}}$, is generated using the current draw of the parameters. The posterior predictive distribution of replicated data is

$$p(\mathbf{X}_{\text{rep}} \mid \mathbf{X}) = \int p(\mathbf{X}_{\text{rep}} \mid \boldsymbol{\omega}) p(\boldsymbol{\omega} \mid \mathbf{X})\, d\boldsymbol{\omega}, \tag{13}$$

where $p(\boldsymbol{\omega} \mid \mathbf{X}) \propto p(\mathbf{X} \mid \boldsymbol{\omega}) p(\boldsymbol{\omega})$ is the posterior of all parameters in the model and $p(\boldsymbol{\omega})$ is the prior of all model parameters. Next, in each country, a frequency distribution of the sum scores is calculated for the replicated data (Béguin and Glas 2001). The posterior predictive score distribution is computed as the mean of the generated frequency distributions over iterations.

Another useful overall fit statistic is the deviance information criterion (Spiegelhalter et al. 2002). It can be used to compare different models such as models that impose versus relax cross-national constraints on item parameters. The model with the lowest expected deviance has the highest posterior probability.

2.5.2. Fit of Specific Model Components. The fit of specific aspects of the model can be evaluated via a Bayesian residual analysis. A Bayesian residual is defined as the difference between the observed response and the expected response, $r_{ik}^g = X_{ik}^g - \sum_{c=1}^{C} c \cdot \Pr(X_{ik}^g = c)$. In each iteration, Bayesian residuals are computed using the sampled values of the model parameters given the data. The computed sequences of Bayesian residuals can be considered to be draws from their marginal posterior distributions. A discrepancy measure $Q(\mathbf{X})$ is a function of the residuals and can be used in a posterior predictive check. For instance, to check item fit, squared residuals across persons for a particular item can be monitored (i.e., $Q(X) = \sum_{i=1}^{n_g} (r_{ik}^g)^2$). To identify items that cause deviations from unidimensionality in a particular country (i.e., residual local dependencies), the conditional covariance between Bayesian residuals concerning two items given the person parameters can be calculated. That is, let $r_k(X, \xi)$ denote the vector of residuals for item k given the data and the person parameters. Subsequently, the covariance between two vectors of item residuals is denoted as $\sigma_{k,k'}(X, \boldsymbol{\xi}) = \text{cov}(r_k(X, \boldsymbol{\xi}), r_{k'}(X, \boldsymbol{\xi}))$, which can be computed within each iteration of the MCMC algorithm using the expected a posteriori estimate for the person parameters. The functions based on the residuals would then be used as a discrepancy measure in a posterior predictive check:

$$P(Q(\mathbf{X}_{\text{rep}}) \geq Q(\mathbf{X}) \mid \mathbf{X}) = \int I(Q(\mathbf{X}_{\text{rep}}) \geq Q(\mathbf{X})) p(\mathbf{X}_{\text{rep}} \mid \mathbf{X})\, d\mathbf{X}_{\text{rep}}. \tag{14}$$

2.6. Item Information Functions

Measurement accuracy in IRT is based on the notion of information. The information function $I(\xi)$ is defined as $I(\xi) = -\text{E}[\partial^2 / \partial \xi^2 \ln L(\mathbf{x} \mid \xi)]$. The category information function $I_{kc}^g(\xi)$ is defined as (Samejima 1969)

$$I_{kc}^g(\xi_i^g) = \frac{[\partial \Pr(x_{ik}^g = c)/\partial \xi_i^g]^2}{\Pr(x_{ik} = c)^2} - \frac{\partial^2 \Pr(x_{ik}^g = c)/\partial \xi_i^{g^2}}{\Pr(x_{ik}^g = c)}. \tag{15}$$

These functions can be merged to yield the item information function $I_k(\xi)$:

$$I_k(\xi_i^g) = \sum_{c=1}^{C} I_{kc}^g(\xi_i^g) \Pr(x_{ik}^g = c) = \sum_{c=1}^{C} \frac{[\partial \Pr(x_{ik}^g = c)/\partial \xi_i^g]^2}{\Pr(x_{ik}^g = c)}. \tag{16}$$

For widely spaced thresholds, the resulting information function might be multimodal, whereas closer thresholds produce unimodal information functions. Closer thresholds provide more information over a smaller range of the ability scale than do more widely spaced thresholds. Finally, the amount of information increases as the number of categories increases (Samejima 1969, Ostini and Nering 2005).

The scale information function (SIF) is the sum of the item information functions $I_k(\xi)$. It is a nonlinear function of the latent variable (Fraley et al. 2000). The SIF has various features that we will use in designing optimal short-form scales. First, when measurement precision is not sufficient at certain levels of the latent variable, items can be added that provide the most information around these levels. Second, measurement precision is not the same for each level of

the latent variable. Items are most informative when the average threshold more or less matches a respondent's ξ value. Third, the amount of information for a test is a linear function of the item information functions, which allows for the application of linear programming techniques in the construction of linear test forms (Van der Linden 2005) and computer-adaptive tests (Dodd et al. 1995, Van Rijn et al. 2002). Note the similarity between this use of information in IRT and adaptive conjoint techniques (Toubia et al. 2003, 2004), adaptive idea screening (Toubia and Florès 2007), and with the literature in educational measurement on assigning greater weights around cut scores (Bradlow and Wainer 1998, Ip 2000).

2.7. Optimal Short Forms

The process of selecting items for a short-form scale, subject to various constraints with some target information function for the measure, can be formalized as a combinatorial optimization problem. In the psychometric literature, test construction using optimization is known as optimal test design (OTD) (Van der Linden 2005). OTD requires one to specify a target information function (TIF). TIF is a function $\tau(\xi)$ that provides the goal values for L grid points ξ_l along the ξ scale. For instance, the goal might be to select those items that yield uniformly good measurement with information level I_{unif} along the trait range (Fraley et al. 2000). If the researcher is primarily interested in high measurement precision at low (ξ_{lo}) and high values (ξ_{hi}) of the latent variable, a bimodal target function is most appropriate. The specification of different shapes of the TIF allows much more flexibility than in CTT where only Cronbach's alpha is usually considered (Hambleton et al. 1991).

The researcher needs to specify the *shape* of the TIF as well as the target information *level* attained by the short form. The standard cutoff for Cronbach's alpha is 0.7. Although this cutoff was developed in the context of CTT, it is a generally accepted criterion that we also adopt in our IRT context. Reliability and information level are related as follows:

$$MRI_g = \frac{\sigma_g^2 - 1/\tau^g}{\sigma_g^2}, \tag{17}$$

where σ_g^2 is the variance in country g (see Equation (4)) and τ^g is the required level of information in country g. Based on Equation (17), we can compute what information level τ^g is required to yield short-form scale reliability of, say, at least 0.7.

The TIFs that we consider are smooth functions. Therefore, it holds that if we require a TIF to meet a smooth target $\tau(\xi)$ at one point on the latent scale, neighborhoods approximate the target as well. Specifying only a small number of grid points thus suffices in normal applications. To assemble the scale with an information function that meets a target, we have a multiobjective assembly problem (Van der Linden 2005, Van der Linden and Boekkooi-Timminga 1989, Veldkamp 1999).

Information functions are subject to uncertainty as a result of the uncertainty in the estimation process. Using the posterior means for the information functions based on the MCMC algorithm implies that some of the information functions are overestimated, whereas for other items, the information functions are underestimated. This might have serious consequences. In the test assembly process, items are selected based on the contribution to the test information function. Thus, item selection might capitalize on positive estimation errors if this uncertainty is not taken into account. Because of this, we implemented a robust optimization method (Ben-Tal and Nemirovski 1998, 1999, 2000).

The ultimate goal of robust optimization is to take data uncertainty into account already at the item selection stage to "immunize" resulting tests against this uncertainty. In robust optimization, constraints will be satisfied whatever the realization of the uncertain (i.e., estimated) parameters within a reasonable prescribed uncertainty set. Our robust optimization procedure is formulated as follows:

$$\begin{aligned} &\text{minimize} \sum_{k=1}^{K_g} ITEM_k^g \\ &\text{subject to} \sum_{k=1}^{K_g} I_k^g(\xi_l^g, \zeta) ITEM_k^g \geq \tau_l^g \quad \forall l, \forall \zeta \in \Im, \\ &\qquad ITEM_k^g \in \{0,1\}, \quad k = 1, \ldots, K_g, \end{aligned} \tag{18}$$

where $ITEM_k^g$ is an indicator (1 = item k included in the scale of country g and 0 = item not included in the scale), $I_k^g(\cdot)$ is the information function of item k in country g, ξ_l^g is the lth grid point in country g, τ_l^g the lth information target level in country g, and $\Im$ is the uncertainty set. To define the uncertainty set, credibility intervals of the information functions can be applied. Finally, for a relatively small number of decision variables (i.e., number of items in the pool < 15), it suffices to use a cubic norm instead of a 95% reliability ellipsoid, and linear programming techniques can be applied instead of nonlinear ones.

Additional constraints can be added to the multiobjective assembly problem. For instance, the constraint $\sum_{\{k:\, d_k^g = d\}} ITEM_k^g \leq 1,\ d = 1, \ldots, D^g$ would specify that only one item is selected from each block of items that display excess correlations (it is assumed that there are D^g blocks of items with residual correlations in country g). It could be desirable for a short-form scale not to display residual correlations. When there are content constraints (e.g., a certain number of items of a certain type are needed), this can be accommodated by adding constraints to the optimization

Table 1 Multigroup Polytomous IRT and Multigroup CFA Models

	Accommodates ordinal data	Hierarchical Bayes	Cross-group varying item parameters	Relaxation of measurement invariance	Etic and emic items	Error correlations among items	Model fit	Scale construction (OTD)
May (2006)	Yes	Yes	Yes	No	Yes	No	No	No
Reise et al. (1993)	Yes	No	Yes	No	No	No	Yes	No
de Jong et al. (2007)	Yes	Yes	Yes	Yes	No	No	No	No
Jöreskog (1971)	No	No	Yes	No	No	Yes	Yes	No
Browne (1984)	Yes	No	Yes	No	No	Yes	Yes	No
This paper	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

model. In educational testing, there are strict demands on deviations from the target, but in marketing, overshooting in terms of information at a particular grid point is not an issue because *MRI* of 0.7 is a minimum rather than an optimum.

2.8. Using the Short Form in Future Research

After the model has been calibrated and new data are collected for the "optimal" items, respondents who would answer completely different items can be compared within and across countries. The item-free calibration property of IRT models ensures that respondents can answer different items and still be compared as long as the items have been calibrated on the same latent scale. Because in the first step all items are simultaneously calibrated on the same latent scale, this assumption is satisfied.

Handbooks of marketing scales could start reporting short-form scales of existing marketing instruments along with their discrimination and threshold parameters. Estimating the latent scores can be done via conditional maximum likelihood (see Hambleton and Swaminathan 1985, Chapter 5) or via Bayesian estimation, where the item parameters are simply fixed.

In summary, the procedure starts at T_0 by estimating the hierarchical IRT model and selecting the optimal items. Once the optimal items are identified and their item parameters estimated, if at a later point in time (T_1) new data are collected, researchers need only to collect data for these short-form items. Latent construct scores for the new sample can be estimated using respondents' scores on the optimal items and the item parameter estimates obtained in T_0. Given that IRT parameter estimates are sample invariant, latent construct scores can be compared with latent scores obtained in previous samples within and across countries. This practice is consistent with the *item banking* literature in psychometrics (see Van der Linden 2005).

2.9. Comparison with Other Models

Our model allows for country-specific subsets of correlated items and a different number of items per country, accommodates etic as well as emic items, relaxes the constraint of measurement invariance, allows construction of short-form scales subject to a variety of researcher-imposed criteria, and enables marketing researchers to compare short-form latent scores across samples within and across countries. In Table 1, we contrast our model with previous multigroup polytomous IRT models as well as with the multigroup ordinal confirmatory factor analysis (MGCFA) model (Browne 1984) and the multigroup metric model (Jöreskog 1971). Table 1 is self-explanatory, but we will briefly elaborate on the differences between the present model and three major alternatives, namely, the multigroup IRT model developed by de Jong et al. (2007), and the off-the-shelf, multigroup ordinal and metric CFA models, all of which are included in popular software packages like LISREL and EQS.

Our model contributes over and above the de Jong et al. (2007) model in four important ways. Our procedure allows for emic items while retaining cross-national comparability. Thus, it addresses the emic-etic dilemma that has haunted international researchers in marketing and other social sciences ever since Berry's (1969) seminal article. It can also accommodate data sets that are less well behaved in some countries. It allows assessment of model fit per country. Finally, it contains a procedure for the selection of optimal short-form scales—subject to a flexible set of user-imposed constraints and targets—that yield cross-nationally comparable latent scores even when the items are different across countries.

Our model also differs in important respects from multigroup metric (Jöreskog 1971) and ordinal (Browne 1984) CFA models. Both multigroup CFA models require at least two invariant items to set the metric of the latent construct (Steenkamp and Baumgartner 1998). Often there will be no invariant items. (Indeed, our empirical application below shows that even for a well-established scale, there are no items that are invariant across countries.) Neither model can accommodate mixtures of country-specific and common items. Item characteristics are sample dependent. Therefore, item banking is not possible, which prohibits comparisons across samples within and between countries.

Ordinal CFA has two other limitations. Scalar invariance cannot be tested, as it is inadmissible to

compute means on ordinal data (Green et al. 1988, p. 244). Hence, latent means cannot be compared across countries (Steenkamp and Baumgartner 1998). Moreover, ordinal CFA typically requires samples exceeding 1,000 per country (Flora and Curran 2004)—which is considerably larger than most sample sizes in marketing research. On the other hand, metric CFA assumes the data to be interval scaled. If this assumption is incorrect, this may lead to invalid conclusions regarding measurement invariance (Lubke and Muthén 2004).

Researchers appear to weigh the limitations of the multigroup ordinal CFA more heavily than the limitations unique to its metric counterpart. Consequently, metric CFA is the method of choice in marketing (Steenkamp and Baumgartner 1998) and other social sciences (Raju et al. 2002, Vandenberg and Lance 2000).

3. Simulation Studies

3.1. Mixed Emic-Etic Scales

The purpose of the first simulation study is to assess the ability of our cross-national IRT model to recover parameter estimates in the presence of emic items across two levels of etic versus emic items. Data were generated with 10 countries, 12 items per country, and 1,000 respondents per country. Each item is measured on a five-point Likert scale. In the condition with a low number of emic items, 3 out of 12 items were country-specific, the other 9 items being common across countries. In the condition with a high number of emic items, 8 out of 12 items were emic and the other four items were etic. In neither condition was any of the etic items invariant across countries.

We used 20,000 burn-in iterations and overdispersed starting values. The next 20,000 iterations were used for inference. The discrimination and threshold parameters are recovered accurately. The mean absolute deviation (MAD) for the discrimination parameters is 0.047 (0.048) for three (eight) emic items, and the MAD for the threshold parameters is 0.062 (0.066) for three (eight) emic items.

Table 2 shows the ability of the model to recover true latent means and variances, even in the presence of emic items. MAD for country means is 0.018 (0.022) for three (eight) emic items, which is very small given the range of the latent variable. The within-country standard deviation of the latent scores was also recovered well, MAD being 0.023 (0.027) for three (eight) emic items. Recovery is accurate even when the clear majority of the items are emic, and there is not a single cross-nationally invariant item.

Table 2 Recovery of Latent Country Means and Standard Deviations in Simulation Study with Etic and Emic Items

	Latent mean		Latent std. dev.	
	True value	Estimated value	True value	Estimated value
	3 emic items versus 9 etic items			
Country 1	−0.526	−0.499	0.894	0.945
Country 2	0.977	1.000	0.691	0.708
Country 3	0.559	0.551	1.355	1.366
Country 4	−0.029	−0.010	0.871	0.920
Country 5	1.375	1.371	0.762	0.752
Country 6	0.670	0.671	0.514	0.534
Country 7	−0.629	−0.574	1.247	1.257
Country 8	0.509	0.500	1.375	1.422
Country 9	0.697	0.678	1.333	1.332
Country 10	0.184	0.199	0.800	0.812
	8 emic items versus 4 etic items			
Country 1	−0.262	−0.283	1.266	1.231
Country 2	−1.213	−1.170	1.166	1.212
Country 3	−1.319	−1.292	0.630	0.625
Country 4	0.931	0.916	0.595	0.587
Country 5	0.011	0.017	0.514	0.531
Country 6	−0.645	−0.657	0.788	0.801
Country 7	0.805	0.817	1.316	1.341
Country 8	0.231	0.251	1.485	1.427
Country 9	−0.989	−0.967	0.517	0.517
Country 10	1.339	1.294	1.319	1.258

3.2. Lack of Unidimensionality

In the second simulation study, we examined the ability of our cross-national IRT model to recover parameter estimates for a longer scale (20 items) with some badly behaving items (is especially common among longer scales) using a sample size ($N = 500$ per country) that is more typical for marketing research. In 3 out of 10 countries, the first six and the last five out of the 20 items were given excess correlations via Equations (10) and (11). The method factor variance was set equal to 75% of the latent variable variance.

As expected, in the countries with excess correlations, the posterior predictive check based on the conditional covariance between Bayesian residuals flagged problems and indicated the sources of the lack of fit (the posterior p-values were zero or one). When two testlet factors are included, model fit improves dramatically. Next, our IRT model is estimated with testlet factors for the three "offending" countries. Item parameters, country means, country variances, and method factor variances are appropriately estimated. MAD for the discrimination (threshold) parameters is 0.076 (0.107), which is a bit higher than it was because of the smaller sample size and larger number of items. The 95% credible interval of the testlet variance [0.82, 1.01] includes its true value of 0.96. The true and estimated country means and variances are accurately recovered as well (see Table 3), MAD being 0.042 and 0.027, respectively.

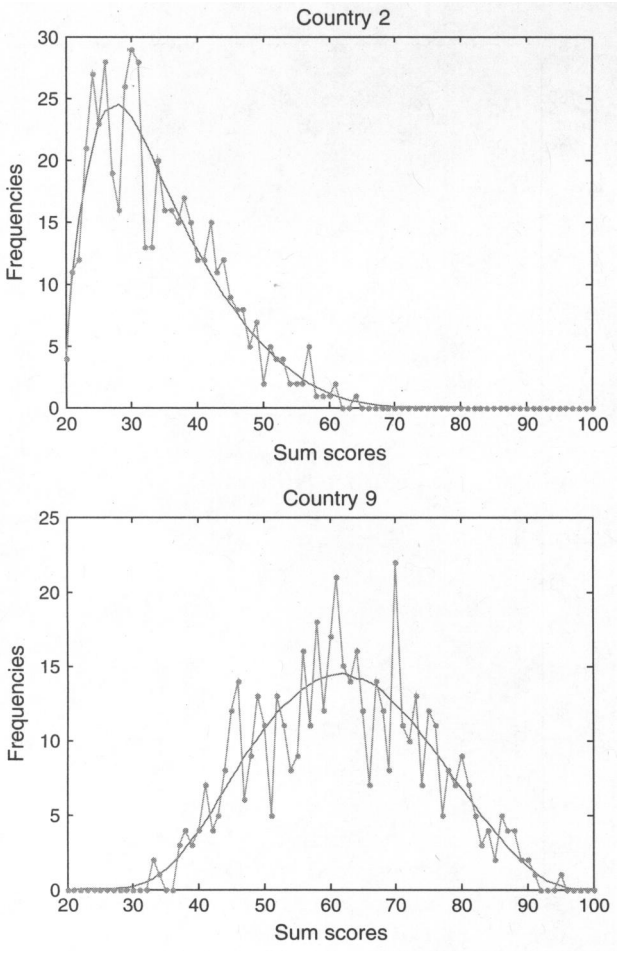
Finally, we examine the fit of our IRT model. In Figure 1, we plot the observed and replicated sum scores for one country with error correlations (country 2) and one country without error correlations (country 9). The

Table 3 Recovery of Latent Country Means and Standard Deviations in Simulation Study with Excess Correlations Between Groups of Items

	Latent mean		Latent std. dev.	
	True value	Estimated value	True value	Estimated value
Country 1	−1.024	−1.028	1.3057	1.274
Country 2	−1.324	−1.380	0.6127	0.651
Country 3	1.057	1.080	1.1487	1.164
Country 4	0.544	0.629	0.9554	0.982
Country 5	−0.928	−0.853	1.3276	1.359
Country 6	1.164	1.129	0.6134	0.588
Country 7	−0.657	−0.720	0.5225	0.522
Country 8	−0.103	−0.134	1.0538	1.089
Country 9	0.418	0.420	0.5387	0.577
Country 10	0.761	0.719	1.1098	1.084

smooth curves represent the mean frequency distributions under our model, whereas the erratic curves are the observed sum scores. Overlap among the curves indicates that the model fits the data well. The same results are observed for other countries.

Figure 1 Observed and Replicated Sum Scores



4. Empirical Application

4.1. Susceptibility to Normative Influence

There is a resurgent interest in social influences on consumer decision making (Bohlmann et al. 2006, Yang and Allenby 2003, Yang et al. 2006). The dominant measure for consumers' susceptibility to normative influences (SNI) is the unidimensional, eight-item scale developed by Bearden et al. (1989). It measures the predisposition to being influenced by others when making purchase decisions. This scale has been used successfully to study social influences on various aspects of consumer behavior such as attitudes toward brands (Batra et al. 2000), consumer confidence (Bearden et al. 1990), protective self-presentation efforts (Wooten and Reed 2004), purchase of new products (Steenkamp and Gielens 2003), and consumer boycotts (Sen et al. 2001). The items of the SNI scale are listed in Table 4.

4.2. Data Collection

Although our model allows for both country-specific as well as common items, data were only collected for the eight original items. Two global marketing research agencies collected data in 28 countries around the world (see Table 5 for the countries). The samples were drawn so as to be broadly representative of the total population in terms of region, age, education, and gender. Some countries used a Web survey, others a mall intercept or hard copy surveys. The number of respondents per country was in the range of 400–600, except for the United States, where the sample size was much larger sample ($n = 1{,}181$). This allows us to do rigorous validation analyses in the United States.

The SNI items were translated into local languages by professional agencies using backtranslation. All items were measured on a five-point Likert scale. We randomly dispersed the SNI items throughout the questionnaire. Bradlow and Fitzsimons (2001)

Table 4 SNI Scale

i1	If I want to be like someone, I often try to buy the same brands that they buy.
i2	It is important that others like the products and brands I buy.
i3	I rarely purchase the latest fashion styles until I am sure my friends approve of them.
i4	I often identify with other people by purchasing the same products and brands they purchase.
i5	When buying products, I generally purchase those brands that I think others will approve of.
i6	I like to know what brands and products make good impressions on others.
i7	If other people can see me using a product, I often purchase the brand they expect me to buy.
i8	I achieve a sense of belonging by purchasing the same products and brands that others purchase.

Table 5 Discrimination Parameters of SNI Scale

	i1	i2	i3	i4	i5	i6	i7	i8
Japan	1.055	0.800	0.706	1.237	1.252	0.733	1.277	1.191
Russia	0.873	1.211	0.410	1.282	1.378	1.208	0.843	1.337
United Kingdom	0.986	0.814	0.552	1.103	1.345	0.985	1.233	1.283
Germany	0.965	0.838	0.444	1.329	1.144	0.981	1.375	1.383
Ireland	0.806	0.743	0.626	1.189	1.175	1.204	1.282	1.257
France	0.835	0.685	0.586	1.284	1.287	1.124	1.507	1.094
Austria	0.881	0.907	0.443	1.272	1.250	0.931	1.408	1.388
The Netherlands	0.833	0.800	0.548	1.227	1.398	1.053	1.189	1.295
Belgium	0.779	0.855	0.499	1.267	1.328	1.224	1.336	1.117
Italy	0.821	0.785	0.493	1.326	1.237	1.108	1.484	1.193
Norway	0.794	0.907	0.647	1.133	1.263	1.046	1.252	1.164
Slovakia	0.938	0.874	0.598	1.326	0.887	0.896	1.326	1.505
Poland	0.942	0.876	0.286	1.245	1.407	1.099	1.544	1.495
Sweden	0.875	0.861	0.620	1.184	1.519	0.993	1.131	1.085
Denmark	0.682	0.900	0.693	1.353	1.264	0.946	1.238	1.193
Hungary	0.851	0.745	0.621	1.114	1.125	1.138	1.409	1.285
Romania	1.213	0.916	0.510	1.201	1.286	0.898	1.251	1.051
United States	0.781	0.802	0.593	1.133	1.342	1.043	1.302	1.314
Argentina	0.871	0.879	0.573	1.006	1.199	1.073	1.450	1.246
Portugal	0.902	0.697	0.680	1.099	1.298	1.095	1.328	1.151
Switzerland	1.021	0.900	0.508	1.393	1.093	0.939	1.341	1.145
Czech Rep.	0.958	0.840	0.517	1.171	1.221	0.890	1.382	1.403
Taiwan	0.938	0.949	0.575	1.321	1.277	0.569	1.557	1.370
Ukraine	1.325	0.892	0.673	1.393	0.936	0.601	1.285	1.284
Brazil	0.880	0.821	0.517	1.284	1.197	1.092	1.223	1.345
Thailand	0.896	0.739	0.845	0.772	1.476	1.186	1.306	1.045
China	1.125	1.064	0.635	1.273	1.042	0.725	1.285	1.102
Spain	0.864	0.709	0.888	0.963	1.326	0.890	1.258	1.307

have noted that randomization may reduce reliability. However, low reliability was not an issue in our study because in all countries, Cronbach's alpha exceeded 0.70.

5. Results

5.1. Estimation

We estimated the hierarchical IRT model and examined both overall fit and the fit of specific aspects of the model. To assess unidimensionality of the SNI scale, we used the posterior predictive check based on the conditional covariance between Bayesian residuals concerning two items given the person parameters. In Russia and Japan, the posterior p-values indicated that there is a large residual error correlation between items 1 and 4 (both posterior p-values equal 1.0). No such lack of fit was identified for the other countries.

Next, we reestimated the hierarchical IRT model, specifying a testlet parameter between items 1 and 4 for Russia and Japan via Equation (10). In both countries, the variance component for the testlet factor is highly significant, validating that the excess correlation identified by posterior predictive checks indeed needs to be included in the IRT model. In Japan, the posterior mean testlet variance is 0.614 (standard deviation = 0.087), whereas in Russia, it is 0.459 (standard deviation = 0.091). As an additional—omnibus—test across all 28 countries, the model with and without the two testlet parameters can be compared using the deviance information criterion (Spiegelhalter et al. 2002). The deviance information criterion for the model without a testlet both for Japan and Russia is 217,645 versus a deviance information criterion of 217,467 for the less restrictive model. The IRT model achieved good overall fit in all countries. For illustrative purposes, Figure 2 shows the plot for one country where the SNI scale was clearly unidimensional (United States) and for a country where items 1 and 4 had excess correlation (Russia).

Table 5 presents the posterior means of the discrimination parameters. There is substantial variation in the discrimination parameters (posterior mean standard deviation is 0.185) across countries. The deviance information criterion of the model is 217,467 compared to 240,551 for a model with invariant discrimination parameters. The variation in discrimination parameters indicates that items measuring a construct well in one country are not always

Figure 2 Observed and Replicated Sum Scores SNI Scale

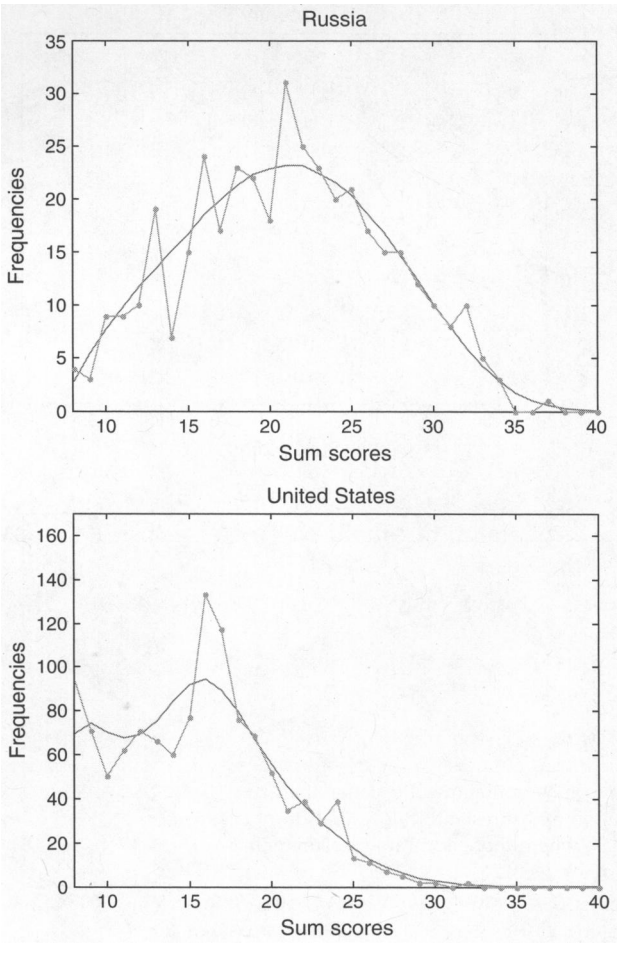
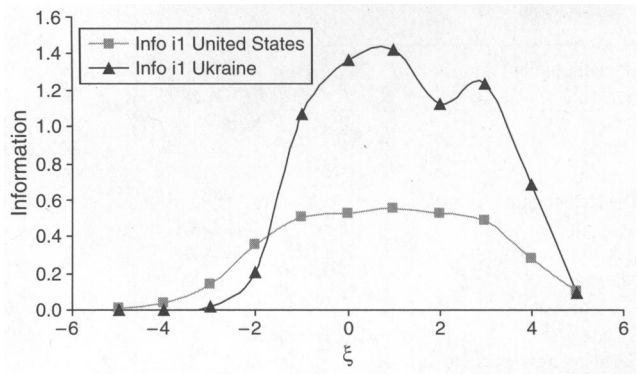


Figure 3 Information Functions Item i1: United States vs. Ukraine



useful in other countries.[2] To illustrate this, consider the United States and the Ukraine. The estimated discrimination parameter of item i1 has a posterior mean of 0.781 (1.325) in the United States (Ukraine). Thus, although the item seems suited to measure SNI in the Ukraine, its posterior mean in the United States is relatively low to be a useful item. When we take the thresholds into account, this is clearly visible in the item information functions. For the United States, the posterior means of the thresholds are $\gamma_{US,1} = -1.225$, $\gamma_{US,2} = 0.420$, $\gamma_{US,3} = 1.391$, and $\gamma_{US,4} = 2.938$, whereas in the Ukraine, the posterior values are $\gamma_{UKR,1} = -0.611$, $\gamma_{UKR,2} = 0.519$, $\gamma_{UKR,3} = 1.532$, and $\gamma_{UKR,4} = 3.262$.

Together with the discrimination parameters, this yields the posterior mean information functions for item i1 in the United States and the Ukraine plotted in Figure 3. The figure shows that (1) measurement precision varies along the latent scale, (2) the information function for the United States is much flatter because of the lower discrimination parameter, and (3) along most of the trait range, the information conveyed by item i1 is much higher in the Ukraine.

The $28 \times 8 \times 4 = 896$ threshold parameters cannot be meaningfully displayed for all other items and countries. The findings of varying thresholds are robust across items. Concordant with the observation of fluctuation in discrimination parameters, there is substantial variation in threshold parameters (posterior mean standard deviation is 0.314).

Table 6 reports the country means and variances, ranked from low to high.[3] China and Taiwan score the highest on SNI. These countries rate high on cultural collectivism, which emphasizes interdependent selves and the importance of social relations (Hofstede 2001). Also, the latent variance in SNI is low in these countries, indicating a relatively high degree of homogeneity with respect to susceptibility to normative influences.

Table 6 Latent Country Means and Variances for SNI (Sorted by Mean)

	Latent mean	Latent variance
The Netherlands	−1.626	1.549
Sweden	−1.477	1.344
Belgium	−1.377	1.269
Denmark	−1.310	1.525
Argentina	−1.232	0.965
United States	−1.215	1.366
Norway	−1.190	1.296
Hungary	−1.173	1.055
Austria	−1.170	1.419
Italy	−1.132	1.292
United Kingdom	−1.046	1.106
France	−1.009	0.980
Spain	−0.994	1.390
Switzerland	−0.952	1.279
Ireland	−0.951	1.248
Portugal	−0.939	1.156
Germany	−0.917	1.194
Czech Rep.	−0.511	0.691
Slovakia	−0.445	0.644
Thailand	−0.410	0.627
Romania	−0.236	0.579
Japan	−0.228	0.716
Poland	−0.195	0.700
Russia	0.010	0.620
Brazil	0.139	0.626
Ukraine	0.318	0.768
Taiwan	0.511	0.391
China	0.932	0.392

The United States and several European countries rate lowest on SNI. People from these countries take the opinions of others on average relatively less into account when making purchase decisions. This may explain why social influences have been relatively understudied in marketing science. After all, most academic research is done in the United States and Western Europe (Stremersch and Verhoef 2005), and social influences play a relatively more modest role in these areas of the world. However, the large latent variance in these countries reveals considerable heterogeneity around the (low) mean. Even in these countries, there are segments of consumers that are clearly susceptible to normative influences, as evidenced by recent work by Bohlmann et al. (2006), Yang and Allenby (2003), and Yang et al. (2006).

5.2. Comparison with the Benchmark Metric CFA Model

It is useful to compare the results of our model to the benchmark multigroup metric CFA model. After all, it is the model of choice in marketing and other social

[2] The threshold parameters are also important when considering the usefulness of items. An item with moderate discrimination parameters may sometimes be better than an item with high discrimination and threshold values that do not match the position of the scale where more accurate measurement is required (see Lord and Novick 1968).

[3] Note that the negative numbers primarily have to do with the scaling of the latent variable via the item parameters.

sciences. Initial estimation of the configural invariance model specifying the same factor structure in all 28 countries revealed a large excess correlation between items 1 and 4 in Russia and Japan. This replicates the findings obtained with the posterior predictive checks for our IRT model. We reestimated the model, specifying correlated errors between items 1 and 4 in these two countries. The overall fit for this configural model is good: $\chi^2(558) = 1{,}965.4$, RMSEA = 0.074, CFI = 0.978, and TLI = 0.968, where RMSEA is the root mean square error of approximation, CFI is the comparative fit index, and TLI is the Tucker-Lewis index. Moreover, all factor loadings are significant and substantial in all countries. Thus, configural invariance is supported (Steenkamp and Baumgartner 1998).

However, substantive interest typically focuses on cross-national comparisons of latent means and variances, which require metric and scalar invariance (Steenkamp and Baumgartner 1998). Imposing metric and scalar invariance leads to a significant decrease in fit: $\Delta\chi^2(378) = 3{,}251.2$, $p < 0.001$. The alternative fit indices also deteriorated substantially. Importantly, the decline in CFI was −0.046. In an extensive simulation study, Cheung and Rensvold (2002) found that CFI is among the most powerful fit indices to distinguish between valid and invalid cross-group constraints. They concluded that if CFI declines by more than 0.01, the null hypothesis of invariance should be rejected (Cheung and Rensvold 2002, p. 251). Furthermore, RMSEA and TLI, which take both model fit and parsimony into account, also deteriorated substantially rather than improved: ΔRMSEA = 0.028, and ΔTLI = −0.025.

A close inspection of the CFA results indicates that the modification indices are always large in a number of countries for each item. This implies that not a single item exhibited invariant loadings and intercepts across all countries. Hence, the well-known metric CFA model cannot be used in the current setting to compare latent means and variances across countries. Note that this setting provides a strong test as SNI is among the best-developed and validated marketing scales. If for such a scale substantive comparisons across countries are problematic, this does not bode well for less-established scales.

5.3. Deriving a Country-Specific Short-Form Version of the SNI Scale

Previous work on susceptibility to normative influences suggests that the entire continuum of social influences from low to high is of interest to marketers (Bearden et al. 1989). Consequently, a uniform specification for the TIF is most appropriate (Fraley et al. 2000). Given the cross-national variation in latent construct variance (see Table 6), we should use a different τ for every country to ensure short-form reliability above 0.7 in each country. Per country, three grid points are chosen along the ξ-scale. Next, we run the robust optimization program (Equation (18)) using GAMS 2.50. For Japan and Russia, a constraint is specified that it is not possible to select both items 1 and 4. We present the selected items in Table 7.

Table 7 Selected Items for Country-Specific Short Form of SNI Scale

	i1	i2	i3	i4	i5	i6	i7	i8
United Kingdom				X	X		X	X
Germany					X	X	X	X
Ireland				X		X	X	X
France		X		X	X	X	X	X
Austria					X		X	X
The Netherlands				X	X			X
Belgium					X	X	X	X
Italy					X	X	X	X
Norway					X	X	X	X
Slovakia	X	X		X	X	X	X	X
Poland				X	X		X	X
Sweden		X		X	X	X	X	X
Denmark				X	X		X	
Hungary					X	X	X	X
Romania	X	X		X	X	X	X	X
United States					X		X	X
Argentina	X	X			X	X	X	X
Portugal					X	X	X	X
Switzerland				X	X	X	X	X
Czech Rep.		X		X	X	X	X	X
Taiwan	X	X	X	X	X	X	X	X
Russia	X	X		X	X	X		X
Ukraine	X			X			X	X
Brazil	X			X	X	X	X	X
Thailand	X	X	X	X	X	X	X	X
China	X	X	X	X	X	X	X	X
Spain					X		X	X
Japan	X	X		X	X		X	X

It can be seen that there is variation in item selection across countries, although items i1 through i3 are not often selected. Interestingly, there are no items that are selected in all countries (even though items 5, 7, and 8 come close), which indicates that a particular item is not equally informative across countries. This calls into question the usual procedure to construct (ad hoc) short-form scales in international marketing research by selecting those items that exhibit high factor loadings in the country where the scale has been developed (e.g., Batra et al. 2000, Ter Hofstede et al. 1999). Using the same short-form items across countries is not optimal, but this procedure is understandable as no short-form scale construction method to date has been able to calibrate items on the same latent scale.

Another interesting finding is that the number of items differs across countries. Thus, the required measurement precision is reached more easily in some countries than in others. For example, all items are required in China and Taiwan, whereas much fewer

items are necessary for most other countries. The reason is that Asian countries have a low latent variable variance (see Table 6), so it is harder to discriminate among respondents in these countries. In such cases, items are needed that discriminate between respondents in the latent zone where almost all respondents are located.

5.4. Validation of Short-Form Scale

A question that arises in the construction of short-form scales is how well it approximates the latent scores based on the full scale. In examining this issue, we focus on the United States, where we have a much larger sample ($N = 1{,}181$). We perform the following steps:

Step 1. Estimate the model based on a random subset of 50% of the respondents and simultaneously derive the optimal short-form scale based on this estimation sample.

Step 2. Compute latent construct scores for the respondents in the holdout sample, using their scores on the short-form items and model parameters as obtained in the estimation sample.

Step 3. Estimate the model based on the holdout sample using all items.

Step 4. Correlate the short-form scores (Step 2) with the full-scale scores (Step 3) and compute the MAD between both sets of scores.

The posterior means of the discrimination parameters for items i1 to i8 in the estimation sample are 0.815, 0.880, 0.580, 1.107, 1.352, 1.011, 1.259, and 1.306. They are very similar to the discrimination parameters for the United States reported in Table 5, the MAD being 0.031. For the short-form scale, we select the same three items as for the full U.S. sample. This short form is used to score the individuals in the holdout sample. The correlation between short-form scores and full-scale scores in the holdout sample is 0.947 ($p < 0.001$). Hence, deleting over 60% of the items leads to a loss of only 10% in information. The MAD of the latent scores is 0.270, which is encouraging given that the total scale range is 2.28.

We conducted additional validation analyses in two other countries with relatively larger samples—Germany ($N = 640$) and Ireland ($N = 552$). The smaller sample sizes in Germany and Ireland required us to use the full sample rather than the holdout sample in Step 3. In both countries, 400 randomly drawn observations were used for model calibration, and the other respondents served as the holdout sample. In Germany, the short-form scale consists of four items. Here, the correlation between the two sets of latent scores is 0.961 ($p < 0.001$), whereas the MAD of the latent scores is 0.216. In Ireland, we obtained a four-item scale, a correlation of 0.952 ($p < 0.001$), and a MAD of 0.224. Thus, the short form approximates the latent scores based on the full scale well, even for smaller samples.

5.5. Validation for Simulated Emic-Etic Version of SNI Scale

A limitation of our empirical study is that we did not have emic items. Although in §3.1 we showed that our procedure can accurately recover parameter estimates in the presence of emic items, it is also useful to assess model performance for combined emic-etic scales with real data. To address this issue, we conducted two validation studies in which we simulated the combined emic-etic condition. It is possible to construct an unbalanced data set with emic and etic items if we purposefully delete some items in particular countries. The key question is whether we can properly estimate the parameters of emic items, which are *not* available in all countries. Note that in the subsequent optimization procedure, it does not matter whether an item is etic or emic. The only thing that matters in country-specific optimization is the marginal posterior distribution of the item information function.

5.5.1. Validation with Simulated Emic Items for the United States. We designated items i5 and i8 as specific to the United States, whereas the other six items are common across all countries. This mimics the condition that not all U.S.-developed items may be relevant in other countries. Thus, a data set is constructed with six etic items in all countries, except for the United States, where two emic items are included as well. We estimate a model based on this data set and compare the item parameters for the emic items in the United States with the item parameters from §5.1. The posterior mean discrimination parameters in Table 5 were 1.342 and 1.314. With the unbalanced data set, the posterior mean estimates are very close: 1.322 and 1.319, respectively. Similarly, the eight estimated threshold parameters resemble those from §5.1, the MAD being only 0.082. The correlation between short-form scores and full-scale scores in the holdout sample is 0.947 ($p < 0.001$). The correlation between scores based on the simulated emic-etic scale and the original data (§5.1) is 0.979 ($p < 0.001$).

5.5.2. Validation with Simulated Emic Items for the Asia-Pacific Region. Previous research has suggested that the Asia-Pacific region may have specific expressions and items that are unique to this region (e.g., Burgess and Steenkamp 2006, Hofstede 2001). To simulate this condition, we conducted a second validation study. We assumed that six items are common to all countries but that items i2 and i7 are specific to the four Asia-Pacific countries (Japan, China, Thailand, and Taiwan). We estimate a model based on this data set and compare the item parameters for the emic items with the item parameters

from §5.1. We find that the discrimination parameters for the emic items closely resemble the results reported in §5.1. Pooled across the Asian countries and items, the MAD is a low 0.039. Thresholds for the emic items were also close to the results reported earlier, the MAD being 0.085. The correlation between scores based on the simulated emic-etic scale and the original data is 0.955 ($p < 0.001$).

6. General Discussion

In the last few decades, measurement of marketing constructs has improved tremendously. Our discipline has started to systematically catalogue our measurement knowledge in handbooks of marketing scales. However, several important issues remain. Existing scales are often too long for effective administration in nonstudent samples. Commonly used CTT statistics depend on the sample and the set of items considered, which precludes item banking. Existing common practice to select high-loading items for the short form does not allow the researcher to measure particular ranges of the latent construct with more precision, even when called for by theory. International research adds additional complications as rigorous (U.S.-developed) scales may exhibit excess correlations between items, items may not be equally informative in other countries, items may not be invariant, and relevant items tapping into local cultural expressions of the construct in question cannot be incorporated if cross-national comparisons are desired.

To address these issues, we propose a new model based on a combination of two powerful psychometric tools: hierarchical item response theory and optimal test design methods. Our procedure can be used to construct a short-form marketing scale in a single country. It can also be applied to multiple countries where local scales can contain common as well as country-specific items. The IRT item parameters are sample invariant and, hence, can be used to score respondents in new samples on the same underlying scale. This allows comparison of new findings with previous findings, whether obtained in the same country or in other countries where the model has been applied before. This is another step toward generating a rigorous bank of marketing data and findings that is characteristic of science.

By extending existing hierarchical item response theory models and by combining it with optimal test design methods, we developed a procedure that yields country-specific short-form marketing scales, yet maintains cross-national comparability of latent scores. As such, our procedure is an important step to addressing the (in)famous emic-etic dilemma that has haunted international marketing research for decades (Burgess and Steenkamp 2006, Craig and Douglas 2001, Kumar 2000).

The procedure is flexible in the sense that the researcher can specify various constraints on item content, scale length, and measurement precision. Researchers can impose that the scale length is constant or impose a fixed minimum precision across countries. Although the latter constraint is more in line with current marketing practice especially in applied studies, scale length may be of even greater concern. Precision can vary for respondents high or low on the trait under investigation. Our procedure can also be used to adapt standardized scales to specific subcultures *within* a country, which hitherto received the same instrument. Countries like the United States have distinct subcultures, with unique attitudes, values, and behavioral expressions, calling for items that are specific for different subcultures (Benet-Martínez and John 1998).

Our model has several limitations that offer avenues for further research. It assumes that a single substantive construct underlies the items—possibly with ill-behaved items. A number of marketing scales (e.g., SERVQUAL, MARKOR) are truly multidimensional in that they consist of multiple, correlated substantive factors. If the dimensional structure is stable across countries (which can be assessed using standard CFA techniques), one solution would be to apply our method to each factor separately. This is a reasonable procedure when the correlations between the substantive factors are modest in magnitude (our work on scale development and analysis suggests that |0.3| is a reasonable cutoff). If the factors are more highly correlated, a simultaneous procedure is preferable. Statistical procedures should be developed to derive short forms for multidimensional scales while taking into account that a certain subscale precision is required.

An even more difficult situation arises when the dimensional structure is unstable countries, because items of a multidimensional scale load on different factors in different countries or because in some countries the unidimensional scale breaks into multiple (possibly different) substantive dimensions. Received insight holds that if a scale lacks stability of factor structure ("configural invariance"), cross-national comparisons are not possible (Steenkamp and Baumgartner 1998, Vandenberg and Lance 2000). Hence, short-form scales do not make sense either. Future model development might address this situation, possibly by focusing on higher other factors.

IRT assumes that the estimated parameters are not affected by their position in the questionnaire. This assumption is not uncontested (Baker 2001). Consequently, the educational literature has recently started to consider context effects (De Boeck and Wilson 2004), but to the best of our knowledge, there are no articles that show empirically that item parameters are heavily affected. More research is needed on

possible context effects and how to deal with them. We speculate that context effects are less strong if items pertaining to the same construct are grouped together (cf. Bradlow and Fitzsimons 2001). In that case, the local context will be constant across surveys.

To date, all IRT models—including our model—specify a reflective relation between items and the latent construct. Recently, one can witness an increased interest in marketing in formative scales (Diamantopoulos and Winklhofer 2001, Jarvis et al. 2003). The concept of short-form scales is difficult to reconcile with the formative logic, which assumes a census of items. Future research should examine this issue in detail.

Our model identifies and controls for cross-national differences in item functioning, but does not provide insight into what causes differential item functioning across countries. A fruitful area of research is to combine our model outcomes with follow-up (qualitative) research to understand why items function differently across countries. Once we have identified the causes of differential item functioning, these factors might be quantified and added as covariates in an extended version of our model.

Finally, it would be interesting to reduce scale length even more via computer adaptive testing (CAT) (Wainer et al. 2000). Such procedures tailor the items to the exact trait levels of respondents. For instance, in health care, doctors are already administering adaptive scales to patients, and the development of CAT is likely to increase with the advent of more and powerful multimedia technology.

Acknowledgments

This article is based on a portion of the first author's doctoral dissertation. The authors gratefully acknowledge AiMark for providing the data. The authors also thank Aharon Ben-Tal and Dick den Hertog for helpful comments.

Appendix. Estimation Details

We use Bayesian inference for the IRT model, in which we specify the posterior distribution of all model parameters. We use data augmentation (Tanner and Wong 1987) to facilitate estimation. By defining a continuous latent variable $\mathbf{Z}$ that underlies the ordinal responses contained in $\mathbf{X}$, it is easier to sample from the conditional distributions of the parameters of interest. Parts of the MCMC algorithm can be found in de Jong et al. (2007). The following steps need to be added:

(1) Sample from $[\psi^g_{i,d} \mid Z, a, \sigma^2_{\psi^g_d}]$, for $d = 1, \ldots, D^g$, where D^g is the total number of subsets of items. Consider subset d and let d^g_k denote the subset of item k in country g. The full conditional distribution for $\psi^g_{i,d}$ is normal with parameters

$$E\left(\psi^g_{i,d} \mid Z, a, \sigma^2_{\psi^g_d}\right) = \frac{\sum_{\{k:\, d^g_k = d\}} a^g_k \left(a^g_k \xi^g_i - Z^g_{ik}\right)}{\sum_{\{k:\, d^g_k = d\}} (a^g_k)^2 + 1/\sigma^2_{\psi^g_d}},$$

$$V\left(\psi^g_{i,d} \mid Z, a, \sigma^2_{\psi^g_d}\right) = \frac{1}{\sum_{\{k:\, d^g_k = d\}} (a^g_k)^2 + 1/\sigma^2_{\psi^g_d}}.$$

(2) Sample from $[a^g_k \mid \xi^g, \mathbf{Z}^g_k, a_k, \sigma^2_a]$, $g = 1, \ldots, G$, $k = 1, \ldots, K_g$.

The prior is $\log a^g_k \sim N(\log a_k, \sigma^2_a)$ for $k \in \Xi$ and $\log a^g_k \sim N(\mu_a, V_a)$ for $k \in \Theta_g$.

The full conditional distribution is the product of the prior and the likelihood. A Metropolis-Hastings algorithm must be used to obtain the samples because the posterior distribution of the item parameters is not a standard distribution.

For identification, it is imposed that $\prod_{k=1}^{K_g} a^g_k = 1$.

(3) Sample from $[\boldsymbol{\gamma}^g_k \mid \boldsymbol{\gamma}_k, \sigma^2_\gamma, a^g_k, Z^g_{ik}, X^g_{ik}]$, $g = 1, \ldots, G$, $k = 1, \ldots, K_g$, and $c = 1, \ldots, C-1$.

The full conditional posterior of the threshold parameters is proportional to

$$\prod_{i \mid g} P\left(\gamma^g_{k,x^g_{ik}} > Z^g_{ik} > \gamma^g_{k,x^g_{ik}-1} \mid \xi^g_i, a^g_k, \boldsymbol{\gamma}^g_k\right) f\left(\boldsymbol{\gamma}^g_k \mid \boldsymbol{\gamma}_k, \sigma^2_\gamma\right) \quad \text{for } k \in \Xi,$$

$$\prod_{i \mid g} P\left(\gamma^g_{k,x^g_{ik}} > Z^g_{ik} > \gamma^g_{k,x^g_{ik}-1} \mid \xi^g_i, a^g_k, \boldsymbol{\gamma}^g_k\right) \quad \text{for } k \in \Theta_g.$$

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 $\gamma^{g,*}_k$ from

$$\gamma^{g,*}_{k,c} \sim N\left(\gamma^{g,m-1}_{k,c}, \sigma^2_{MH}\right) I\left(\gamma^{g,*}_{k,c-1} < \gamma^{g,*}_{k,c} < \gamma^{g,m-1}_{k,c+1}\right) \quad \text{for } c = 1, \ldots, C-1,$$

where σ^2_{MH} is a tuning parameter to adjust the accept/reject rate of the algorithm. The Metropolis-Hastings acceptance probability is then given by

$$\min\left[\prod_{i \mid g} \frac{\Pr\left(X^g_{ik} = x^g_{ik} \mid \xi^g_i, a^g_k, \boldsymbol{\gamma}^{g,*}_k\right)}{\Pr\left(X^g_{ik} = x^g_{ik} \mid \xi^g_i, a^g_k, \boldsymbol{\gamma}^{g,m-1}_k\right)} \cdot \frac{f\left(\boldsymbol{\gamma}^{g,*}_k \mid \boldsymbol{\gamma}_k, \sigma^2_{\gamma_k}\right)}{f\left(\boldsymbol{\gamma}^{g,m-1}_k \mid \boldsymbol{\gamma}_k, \sigma^2_{\gamma_k}\right)} \frac{f\left(\boldsymbol{\gamma}^{g,m-1}_k \mid \boldsymbol{\gamma}^{g,*}_k, \sigma^2_{MH}\right)}{f\left(\boldsymbol{\gamma}^{g,*}_k \mid \boldsymbol{\gamma}^{g,m-1}_k, \sigma^2_{MH}\right)}, 1\right] \quad \text{for } k \in \Xi,$$

and

$$\min\left[\prod_{i \mid g} \frac{\Pr\left(X^g_{ik} = x^g_{ik} \mid \xi^g_i, a^g_k, \boldsymbol{\gamma}^{g,*}_k\right)}{\Pr\left(X^g_{ik} = x^g_{ik} \mid \xi^g_i, a^g_k, \boldsymbol{\gamma}^{g,m-1}_k\right)} \cdot \frac{f\left(\boldsymbol{\gamma}^{g,m-1}_k \mid \boldsymbol{\gamma}^{g,*}_k, \sigma^2_{MH}\right)}{f\left(\boldsymbol{\gamma}^{g,*}_k \mid \boldsymbol{\gamma}^{g,m-1}_k, \sigma^2_{MH}\right)}, 1\right] \quad \text{for } k \in \Theta_g.$$

The first part of the expressions represent the contributions from the likelihood, and the second parts come from the proposal distributions. For identification, we set $\sum_{k=1}^{K_g} \gamma^g_{k3} = 0$.

(4) $[\sigma^2_{\psi^g_d} \mid rest]$.

For the conditional distributions, an inverse gamma prior is specified with parameters g_1 and g_2. As a result, each full conditional has an inverse gamma distribution with shape parameter $N_g/2 + g_1$, respectively, and scale parameter $g_2 + \sum_{i=1}^{N_g} (\psi^g_{i,d})^2/2$. Noninformative proper priors were specified with $g_1 = g_2 = 1$.

References

Aaker, J. L., V. Benet-Martínez, J. Garolera. 2001. Consumption symbols as carriers of culture: A study of Japanese and Spanish brand personality constructs. *J. Personality Soc. Psych.* **81**(3) 492–508.

Baker, F. B. 2001. *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park.

Batra, R., V. Ramaswamy, D. L. Alden, J.-B. E. M. Steenkamp, S. Ramachander. 2000. Effects of brand local and nonlocal origin on consumer attitudes in developing countries. *J. Consumer Psych.* **9**(2) 83–95.

Baumgartner, H. 2004. Issues in assessing measurement invariance in cross-national research. Presentation, Sheth Foundation/Sudman Symposium Cross-Cultural Survey Research, University of Illinois, Urbana.

Baumgartner, H., J.-B. E. M. Steenkamp. 2006. An extended paradigm for measurement analysis applicable to panel data. *J. Marketing Res.* **43**(August) 431–442.

Bearden, W. O., R. G. Netemeyer. 1999. *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*, 2nd ed. Sage Publications, Newbury Park, CA.

Bearden, W. O., R. G. Netemeyer, J. E. Teel. 1989. Measurement of consumer susceptibility to interpersonal influence. *J. Consumer Res.* **15**(March) 473–481.

Bearden, W. O., R. G. Netemeyer, J. E. Teel. 1990. Further validation of the consumer susceptibility to interpersonal influence scale. M. E. Goldberg, G. Gorn, R. W. Pollay, eds. *Advances in Consumer Research*, Vol. 17. Association for Consumer Research, Provo, UT, 770–776.

Béguin, A. A., C. A. W. Glas. 2001. MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika* **66**(4) 541–562.

Benet-Martínez, V., O. P. John. 1998. *Los Cinco Grandes* across cultures and ethnic groups: Multitrait multimethod analyses of the big five in Spanish and English. *J. Personality Soc. Psych.* **75**(3) 729–750.

Ben-Tal, A., A. Nemirovski. 1998. Robust convex optimization. *Math. Oper. Res.* **23**(4) 769–805.

Ben-Tal, A., A. Nemirovski. 1999. Robust solutions to uncertain linear programs. *Oper. Res. Lett.* **25**(1) 1–13.

Ben-Tal, A., A. Nemirovski. 2000. Robust solutions of linear programming problems contaminated with uncertain data. *Math. Programming* **88**(3) 411–424.

Bergkvist, L., J. R. Rossiter. 2007. The predictive validity of multiple-item versus single-item measures of the same constructs. *J. Marketing Res.* **44**(May) 175–184.

Berry, J. W. 1969. On cross-cultural comparability. *Internat. J. Psych.* **4**(2) 119–128.

Bohlmann, J. D., J. A. Rosa, R. N. Bolton, W. D. Qualls. 2006. The effect of group interactions on satisfaction judgments: Satisfaction escalation. *Marketing Sci.* **25**(4) 301–321.

Bollen, K. A. 1989. *Structural Equations with Latent Variables*. John Wiley & Sons, New York.

Bolton, R. N. 2003. From the editor. *J. Marketing* **67**(January) 1–3.

Bradlow, E. T., G. J. Fitzsimons. 2001. Subscale distance and item clustering effects in self-administered surveys: A new metric. *J. Marketing Res.* **38**(May) 254–261.

Bradlow, E. T., H. Wainer. 1998. Some statistical and logical considerations when rescoring tests. *Statistica Sinica* **8** 713–728.

Bradlow, E. T., H. Wainer, X. Wang. 1999. A Bayesian random effects model for testlets. *Psychometrika* **64**(2) 153–168.

Browne, M. W. 1984. Asymptotically distribution-free methods for the analysis of covariance structures. *British J. Math. Statist. Psych.* **37**(1) 62–83.

Bruner, G. C., P. J. Hensel. 1992. *Marketing Scales Handbook: A Compilation of Multi-Item Measures*. American Marketing Association, Chicago.

Burgess, S. M., J.-B. E. M. Steenkamp. 2006. Marketing renaissance: How research in emerging markets advances marketing science and practice. *Internat. J. Res. Marketing* **23**(4) 337–356.

Cheung, G. W., R. B. Rensvold. 2002. Evaluating goodness-of fit indexes for testing measurement invariance. *Structural Equation Modeling* **9**(2) 233–255.

Craig, C. S., S. P. Douglas. 2001. *International Marketing Research*, 2nd ed. John Wiley & Sons, New York.

De Boeck, P., M. Wilson. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer-Verlag, New York.

de Jong, M. G., J.-B. E. M. Steenkamp, J.-P. Fox. 2007. Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *J. Consumer Res.* **34**(August) 260–278.

Diamantopoulos, A., H. M. Winklhofer. 2001. Index construction with formative indicators: An alternative to scale development. *J. Marketing Res.* **38**(May) 269–277.

Dodd, B. G., R. J. De Ayala, W. R. Koch. 1995. Computerized adaptive testing with polytomous items. *Appl. Psych. Measurement* **19**(1) 5–22.

Flora, D. B., P. J. Curran. 2004. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psych. Methods* **9**(4) 466–491.

Fraley, C. R., N. G. Waller, K. A. Brennan. 2000. An item response theory analysis of self-report measures of adult attachment. *J. Personality Soc. Psych.* **78**(2) 350–365.

Gerbing, D. W., J. C. Anderson. 1988. An updated paradigm for scale development incorporating unidimensionality and its assessment. *J. Marketing Res.* **25**(May) 186–192.

Green, P. E., D. S. Tull, G. Albaum. 1988. *Research for Marketing Decisions*, 5th ed. Prentice Hall, Englewood Cliffs, NJ.

Gupta, S., V. A. Zeithaml. 2006. Customer metrics and their impact on financial performance. *Marketing Sci.* **25**(6) 718–739.

Hambleton, R. K., H. Swaminathan. 1985. *Item Response Theory: Principles and Applications*. Kluwer-Nijhof, Boston.

Hambleton, R. K., H. Swaminathan, H. J. Rogers. 1991. *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, CA.

Hofstede, G. 2001. *Culture's Consequences*, 2nd ed. Sage, Thousand Oaks, CA.

Ip, E. H.-S. 2000. Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika* **65**(1) 73–91.

Jarvis, C. B., S. B. MacKenzie, P. M. Podsakoff. 2003. A critical review of construct indicators and measurement model specification in marketing and consumer research. *J. Consumer Res.* **30**(September) 199–218.

Johnson, V. E., J. H. Albert. 1999. *Ordinal Data Modeling*. Springer, New York.

Johnson, M. S., S. Sinharay, E. T. Bradlow. 2006. Hierarchical item response theory. C. R. Rao, S. Sinharay, eds. *Handbook of Statistics*, Vol. 26. Elsevier/North-Holland, Amsterdam, 587–604.

Jöreskog, K. G. 1971. Simultaneous factor analysis in several populations. *Psychometrika* **36**(4) 409–426.

Kumar, V. 2000. *International Marketing Research*. Prentice Hall, Upper Saddle River, NJ.

Lord, F. M., M. R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Reading, MA.

Lubke, G. H., B. O. Muthén. 2004. Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Model.* **11**(4) 514–534.

May, H. 2006. A multilevel Bayesian IRT method for scaling socioeconomic status in international studies of education. *J. Educational Behav. Statist.* **31**(Spring) 63–79.

Netemeyer, R. G., W. O. Bearden, S. C. Sharma. 2003. *Scaling Procedures: Issues and Applications*. Sage, Thousand Oaks, CA.

Ostini, R., M. L. Nering. 2005. *Polytomous Item Response Theory Models*. Sage, Thousand Oaks, CA.

Raju, N. S., B. M. Byrne, L. J. Laffitte. 2002. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *J. Appl. Psych.* **87**(3) 517–529.

Reise, S. P., K. F. Widaman, R. H. Pugh. 1993. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psych. Bull.* **114**(3) 552–566.

Richins, M. L. 2004. The material values scale: A re-inquiry into its measurement properties and the development of a short form. *J. Consumer Res.* **31**(S1) 209–219.

Samejima, F. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* **17** 1–100.

Sen, S., Z. Gürhan-Canli, V. G. Morwitz. 2001. Withholding consumption: A social dilemma perspective on consumer boycotts. *J. Consumer Res.* **28**(December) 399–417.

Shimp, T. A., S. Sharma. 1987. Consumer ethnocentrism: Construction and validation of the CETSCALE. *J. Marketing Res.* **24**(August) 280–289.

Shugan, S. M. 2006. Who is afraid to give freedom of speech to marketing folks? (Editorial.) *Marketing Sci.* **25**(5) 403–410.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, A. van der Linde. 2002. Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc. Ser. B* **64**(October) 583–639.

Steenkamp, J.-B. E. M. 2005. Moving out of the U.S. silo: A call to arms for conducting international marketing research. *J. Marketing* **69**(October) 6–8.

Steenkamp, J.-B. E. M., H. Baumgartner. 1995. Development and cross-cultural validation of a short form of CSI as a measure of optimum stimulation level. *Internat. J. Res. Marketing* **12**(2) 97–104.

Steenkamp, J.-B. E. M., H. Baumgartner. 1998. Assessing measurement invariance in cross-national consumer research. *J. Consumer Res.* **25**(June) 78–90.

Steenkamp, J.-B. E. M., K. Gielens. 2003. Consumer and market drivers of the trial rate of new consumer products. *J. Consumer Res.* **30**(December) 368–384.

Stremersch, S., P. C. Verhoef. 2005. Globalization of authorship in the marketing discipline: Does it help or hinder the field? *Marketing Sci.* **24**(4) 585–594.

Tanner, M. A., W. H. Wong. 1987. The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**(398) 528–550.

Ter Hofstede, F., J.-B. E. M. Steenkamp, M. Wedel. 1999. International market segmentation based on consumer-product relations. *J. Marketing Res.* **36**(February) 1–17.

Thompson, E. R. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *J. Cross-Cultural Psych.* **38**(March) 227–242.

Toubia, O., L. Florès. 2007. Adaptive idea screening using consumers. *Marketing Sci.* **26**(3) 342–360.

Toubia, O., J. R. Hauser, D. I. Simester. 2004. Polyhedral methods for adaptive choice-based conjoint analysis. *J. Marketing Res.* **41**(February) 116.

Toubia, O., D. I. Simester, J. R. Hauser, E. Dahan. 2003. Fast polyhedral adaptive conjoint analysis. *Marketing Sci.* **22**(3) 273–303.

Tuerlinckx, F., P. De Boeck. 2004. Models for residual dependencies. P. De Boeck, M. Wilson, eds. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* Springer-Verlag, New York, 289–316.

Vandenberg, R. J., C. E. Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* **3**(1) 4–70.

Van der Linden, W. J. 2005. *Linear Models for Optimal Test Design.* Springer, New York.

Van der Linden, W. J., E. Boekkooi-Timminga. 1989. A maximin model for test design with practical constraints. *Psychometrika* **54**(2) 237–247.

Van Rijn, P. W., T. J. H. M. Eggen, B. T. Hemker, P. F. Sanders. 2002. Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Appl. Psych. Measurement* **26**(4) 393–411.

Veldkamp, B. P. 1999. Multiple objective test assembly problems. *J. Educational Measurement* **36**(Autumn) 253–266.

Wainer, H., E. T. Bradlow, Z. Du. 2000. Testlet response theory: An analog for the 3-PL useful in testlet-based adaptive testing. W. J. Van der Linden, C. A. W. Glas, eds. *Computerized Adaptive Testing, Theory and Practice.* Kluwer-Nijhof, Boston, 245–270.

Wang, X., E. T. Bradlow, H. Wainer. 2002. A general Bayesian model for testlets: Theory and applications. *Appl. Psych. Measurement* **26**(March) 109–128.

Winer, R. S. 1998. From the editor. *J. Marketing Res.* **35**(February) iii–iv.

Wooten, D. B., A. Reed. 2004. Playing it safe: Susceptibility to normative influence and protective self-presentation. *J. Consumer Res.* **31**(December) 551–556.

Yang, S., G. M. Allenby. 2003. Modeling interdependent consumer preferences. *J. Marketing Res.* **40**(August) 282–294.

Yang, S., V. Narayan, H. Assael. 2006. Estimating the interdependence of television program viewership between spouses: A Bayesian simultaneous equation model. *Marketing Sci.* **25**(4) 336–349.