MARTIJN G. DE JONG, JAN-BENEDICT E.M. STEENKAMP, JEAN-PAUL FOX, and HANS BAUMGARTNER*

Extreme response style (ERS) is an important threat to the validity of survey-based marketing research. In this article, the authors present a new item response theory–based model for measuring ERS. This model contributes to the ERS literature in two ways. First, the method improves on existing procedures by allowing different items to be differentially useful for measuring ERS and by accommodating the possibility that an item's usefulness differs across groups (e.g., countries). Second, the model integrates an advanced item response theory measurement model with a structural hierarchical model for studying antecedents of ERS. The authors simultaneously estimate a person's ERS score and individual- and group-level (country) drivers of ERS. Through simulations, they show that the new method improves on traditional procedures. They further apply the model to a large data set consisting of 12,506 consumers from 26 countries on four continents. The findings show that the model extensions are necessary to model the data adequately. Finally, they report substantive results about the effects of socio-demographic and national-cultural variables on ERS.

Keywords: item response theory, response styles, scale usage, testlets, systematic measurement error, varying item parameters, measurement invariance, international marketing research

Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation

Valid measurement is a cornerstone of marketing as a science. Although the measurement of marketing constructs has greatly improved in recent years, systematic error is often neglected. However, it is well known that responses to questionnaires are often influenced by content-irrelevant factors called "response styles" (Baumgartner and Steenkamp 2001, 2006). A response style can be defined as a person's tendency to respond systematically to questionnaire items on some basis other than what the items were specifically designed to measure (Paulhus 1991).

In this article, we focus on extreme response style (ERS), one of the most pervasive and frequently studied response styles in the social sciences (see, e.g., Baumgartner and Steenkamp 2001; Greenleaf 1992b; Johnson 2003; Paulhus 1991). Extreme response style is the tendency of respondents to favor or avoid using the endpoints of a rating scale, relatively independently of specific item content. Although the literature on ERS is extensive, the phenomenon has received relatively little attention in marketing journals (cf. Baumgartner and Steenkamp 2001; Greenleaf 1992a). This is surprising because ERS has biasing effects on both the mean level of responses and the correlation between marketing constructs (Baumgartner and Steenkamp 2001; Greenleaf 1992a; Rossi, Gilula, and Allenby 2001). Furthermore, in cross-national marketing research, country-

*Martijn G. de Jong is Assistant Professor of Marketing, RSM Erasmus University (e-mail: MJong@rsm.nl). Jan-Benedict E.M. Steenkamp is C. Knox Massey Distinguished Professor of Marketing and Marketing Area Chair, Kenan-Flagler Business School, University of North Carolina at Chapel Hill (e-mail: JBS@unc.edu). Jean-Paul Fox is an assistant professor, Department of Research Methodology, Measurement, and Data Analysis, University of Twente (e-mail: G.J.A.Fox@edte.utwente.nl). Hans Baumgartner is Professor of Marketing and Charles & Lillian Binder Faculty Fellow, Smeal College of Business, Pennsylvania State University (e-mail: jxb14@psu.edu). This article is based on the first author's doctoral dissertation, written when he was a doctoral student at Tilburg University. The authors thank the anonymous *JMR* reviewers for their extremely useful and constructive comments. In addition, they thank AiMark for providing the data and gratefully acknowledge financial support from the Flemish Science Foundation (Grant No. G.0116.04).

ISSN: 0022-2437 (print), 1547-7193 (electronic)

Journal of Marketing Research
Vol. XLV (February 2008), 104–115

specific variations in ERS may easily be misinterpreted as substantive differences in the marketing constructs examined, which could have adverse effects on international marketers' decisions (Kumar 2000). Thus, ERS is an important threat to the validity of both domestic and cross-national survey-based marketing research.

The current research contributes to the ERS literature in two ways. First, we propose a new method based on item response theory (IRT) for measuring ERS. Our method improves on existing procedures (Baumgartner and Steenkamp 2001; Greenleaf 1992b) by allowing different items to be differentially useful for measuring ERS and by accommodating the possibility that an item's usefulness differs across groups (e.g., countries). Furthermore, the method relaxes the requirement that the items in an ERS measure should be (marginally) uncorrelated (Greenleaf 1992b), which allows marketing researchers to construct an ERS measure based on substantively correlated items and eliminates the need for a dedicated ERS scale. Through simulations, we show that the new method improves on traditional procedures, and a detailed analysis of a large-scale data set indicates that the modifications are necessary to model the data adequately. The ERS score can subsequently be used to correct survey data for ERS bias.

Second, our model integrates the advanced IRT measurement model with a structural hierarchical model for studying the antecedents of ERS. We simultaneously estimate a person's ERS score and individual- and group-level (country) drivers of ERS, thus providing insights into the determinants of this important response style across people and countries. Specifically, we study both sociodemographic and national-cultural determinants of ERS using a data set consisting of 12,500 consumers from 26 countries on four continents.

MEASURING ERS

Traditional Approaches to ERS Measurement

The observed score on any marketing scale X can be partitioned into three components:

$$X_i = T_i + S_i + E_i, \tag{1}$$

where T_i is the latent true score of respondent i, S_i is systematic error, and E_i is random error. An important cause of systematic error is ERS (Greenleaf 1992b). To purge ERS from construct measurements, marketing researchers have proposed partialing systematic influences due to ERS from scale scores with a three-step procedure: (1) construct an estimate of a person's ERS score based on a set of items, (2) regress observed scores for substantive scales on ERS, and (3) use the purified scores in further analyses (Baumgartner and Steenkamp 2001).

An ERS estimate is typically constructed by summing the number of extreme responses that a respondent endorses across a set of items (Baumgartner and Steenkamp 2001). For example, with a five-point Likert scale, an ERS measure corresponds to the number of questionnaire statements with which a respondent "strongly agrees" or "strongly disagrees" (Greenleaf 1992b). We can display this as

$$\hat{ERS}_i = \sum_{k=1}^{K} I_{\{Q_{ik}=1 \vee Q_{ik}=5\}} = \sum_{k=1}^{K} EXTR_{ik}, \tag{2}$$

where $EXTR_{ik} = I_{\{Q_{ik}=1 \vee Q_{ik}=5\}}$ is an indicator variable that takes the value of 1 when respondent i uses an extreme response option on the Likert scale for question k (Q_{ik}) and 0 if otherwise, $\hat{ERS}_i$ represents an estimate of the respondent's latent ERS score, and K equals the number of items.

Equation 2 specifies that at the observed level of individual items, ERS is measured on a dichotomous scale. We retain this basic operationalization of ERS measurement in our proposed IRT-based model because of the following reasons: First, this specification is commonly used in the marketing literature and in other social sciences (Bachman and O'Malley 1984; Baumgartner and Steenkamp 2001; Chen, Lee, and Stevenson 1995; Greenleaf 1992b; Grim and Church 1999; Hui and Triandis 1989; Johnson et al. 2005; Marín, Gamba, and Marín 1992). Second, it is an obvious and intuitive operationalization of extreme responding for the five- or seven-point scales most commonly used in marketing survey research (Bearden and Netemeyer 1999). Scaling experts sometimes operationally define ERS in this way. For example, Paulhus (1991, p. 49) notes that ERS is the "tendency to use the extreme choices on a rating scale (e.g., 1s and 7s on a seven-point scale)." Third, the dichotomization minimizes confounding ERS with acquiescence responding, which is often operationalized as follows:

$$\hat{ARS}_i = \sum_{k=1}^{K} (2 \times I_{\{Q_{ik}=5\}} + I_{\{Q_{ik}=4\}}). \tag{3}$$

Because disacquiescence is much less common than acquiescence (Baumgartner and Steenkamp 2001), this implies that an absolute deviation measure of ERS (e.g., by coding a response of 2 or 4 on a five-point scale as 1 and a response of 1 or 5 as 2) will overlap substantially with acquiescence responding. This will be much less the case for a dichotomous ERS measure. This is a major reason researchers have used the dichotomous measure.

Limitations of Traditional Approaches to ERS Measurement

Two approaches for measuring ERS can be distinguished on the basis of which items are included in Equation 2: (1) the use of dedicated ERS instruments and (2) the use of ad hoc measures of ERS based on items intended to assess substantive constructs.[1]

Dedicated ERS instruments. Survey researchers sometimes use a separate set of items that were specifically designed to measure ERS. Although seemingly attractive at first sight, this approach has some significant disadvantages. First, few dedicated ERS scales exist (Greenleaf 1992b is a notable exception). Second, adding nonsubstantive items to a survey is costly in terms of both money and respondent fatigue. It is often difficult to get marketers to pay for additional survey items that will be used solely for estimating stylistic responding. Using items that are already included in the survey can lower the cost and time involved.

[1]We thank an anonymous reviewer for suggesting several arguments in this section.

Third, if the ERS properties of dedicated ERS scale items vary across subgroups (which will probably be the norm), it may be futile for researchers to try to assemble a set of items that will work equally well across cultural and linguistic subgroups. This issue is particularly problematic in international marketing research. Fourth, using proven items from existing substantive scales is advantageous in cross-linguistic research because these are the kinds of items that have been thoroughly tested in many languages, using procedures such as back translation.

Ad hoc ERS items based on substantive scales. For all these reasons, it is common to use items that were originally designed to measure substantive constructs as indicators of ERS (e.g., Baumgartner and Steenkamp 2001; Greenleaf 1992a; Van Herk, Poortinga, and Verhallen 2004). If such ad hoc ERS scales are used to partial stylistic variance from substantive scales, it is critical that there is no item overlap between the ERS measure and the substantive scales that are to be purged of systematic error variance (Baumgartner and Steenkamp 2001).

However, even if item overlap is avoided, the traditional method still has serious limitations. First, as Greenleaf (1992b) notes, items in an ERS measure should have low average interitem correlations for substantive reasons (to avoid confounding style and content). However, when ad hoc measures of ERS are constructed on the basis of substantive scales, items from the same scale will be correlated and, indeed, should be correlated (Bearden and Netemeyer 1999). The traditional ERS formula in Equation 2 ignores this dependence structure.

Second, ERS is best understood as an interaction of personal dispositions and item characteristics (Podsakoff et al. 2003). Respondents differ in their tendency to go to the extremes of the rating scale, and items elicit ERS to differing degrees. Equation 2 does not allow for this, because it does not separate item and person effects but assigns equal weights to all items.

Third, the usefulness of an item for measuring ERS may vary across countries and linguistic subgroups. Cross-national differences in the ERS properties of items may arise because of differences in item semantics and cultural meaning (Podsakoff et al. 2003). Proportions of extreme responses are likely to differ across countries and subgroups, and ERS measures based on different items would be incomparable across countries.

Finally, when survey researchers want to examine individual and national drivers of ERS, it is important to integrate the measurement model for ERS with a structural hierarchical latent variable model and to estimate all parameters simultaneously to avoid bias in parameter estimates (Ansari, Jedidi, and Jagpal 2000; Fox and Glas 2001).[2] What is needed is a model that adjusts for differences in item characteristics across items and subgroups, accounts for substantive correlations among items from the same scale, and allows the researcher to study the antecedents of ERS simultaneously. In the next section, we propose such a model.

[2]Note that the last three limitations also hold for a dedicated ERS measure.

MEASURING ERS USING IRT

The Basic IRT Model

To address the limitations of the existing operationalization of ERS, we use a binary IRT measurement model (Lord and Novick 1968). Binary IRT models are a powerful approach for relating multiple dichotomous observed variables to an underlying continuous latent trait (Hambleton and Swaminathan 1985; Lord and Novick 1968). We assume that a continuous, stable, latent ERS trait underlies a person's observed extreme response pattern (Baumgartner and Steenkamp 2001; Greenleaf 1992b)—that is, the dichotomous pattern of zeros and ones contained in $\mathbf{EXTR}_i = (EXTR_{i1}, \ldots, EXTR_{iK})'$. We follow previous research in assuming that the observed indicators of ERS are measured on a dichotomous scale because the reasons supporting this practice in the context of the traditional ERS measurement apply equally well to the IRT model.[3] However, we adopt a radically different approach for modeling the relationship between the latent ERS construct and its observed indicators.

Item response theory models have a cross-classified character with separate item and person characteristics. Thus, IRT is suited to separating the influence of items (How easily does item k elicit ERS?) and people (What is the latent ERS score of person i?) with respect to an observed extreme response $EXTR_{ik}$. Note that, except for Greenleaf (1992b), researchers have typically assumed that each item is equally useful for measuring ERS. However, it is likely that items differ in their tendency to elicit extreme responses (Bradlow and Zaslavsky 1999; Podsakoff et al. 2003).

A frequently used IRT model is the two-parameter normal ogive model (Lord and Novick 1968). For this model, the probability of an extreme response for respondent i on Likert item k (i.e., "strongly disagree" or "strongly agree" so that $EXTR_{ik} = 1$) is driven by the respondent's latent ERS value, random error, and item characteristics (e.g., specific item content, semantics). Mathematically, the two-parameter normal ogive is formulated as follows:

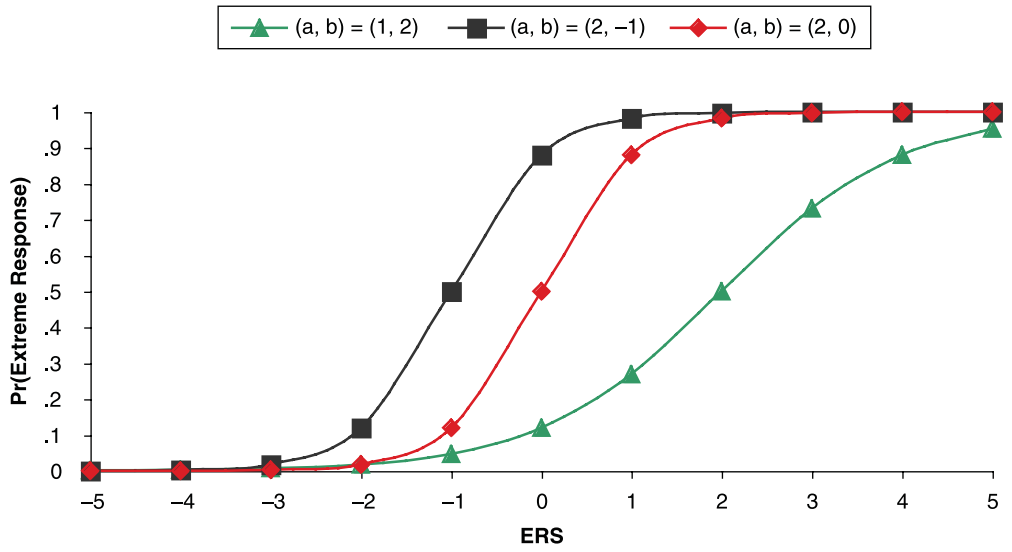
$$(4) \qquad P(EXTR_{ik} = 1|ERS_i, a_k, b_k) = \Phi[a_k(ERS_i - b_k)],$$

where a_k is the discrimination parameter for item k, b_k is the "difficulty" or threshold parameter for item k, and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The function $\Phi[a_k(ERS_i - b_k)]$ is known as the item characteristic curve. Some examples of item characteristic curves for different combinations of a (discrimination) and b (difficulty) appear in Figure 1.

The difficulty parameter b_k, which is measured on the same scale as ERS_i, indicates how likely it is that an item k will elicit an extreme response. Items with a negative b_k parameter elicit an extreme response easily, whereas items

[3]Although for ordinal data a graded-response IRT model (Samejima 1969) is a higher information method for measuring item characteristics, a graded IRT model is not suitable for measuring ERS. This is because the latent trait in the graded IRT model would capture a general method factor (Podsakoff et al. 2003) rather than ERS. The threshold parameters in a graded IRT model would not reflect an item's ERS properties.

Figure 1
EXAMPLES OF ITEM CHARACTERISTIC CURVES



with a positive b_k parameter do not readily evoke an extreme response. Technically, b_k is defined such that a respondent i with $ERS_i = b_k$ has a probability of .5 of making an extreme response on item k. In other words, the parameter b_k determines the inflection point of the S-shaped curve (see Figure 1).

The discrimination parameter a_k determines whether an item discriminates well between people who are high on ERS and those who are low. It is conceptually similar to a factor loading in confirmatory factor analysis because it represents the relationship between the latent ERS score and observed item responses. Items with an a_k value close to zero are not useful for measuring ERS. Note that a_k assesses an item's effectiveness as an indicator of ERS, not its substantive validity. For a high value of a_k, an extreme response provides strong evidence that the ERS_i value is above b_k. In Figure 1, the parameter a_k determines the steepness of the curve.

Furthermore, note that the usefulness of an item is largely conditional on the particular location of the trait level ERS_i. If we had an a priori expectation that the ERS value for respondent i was in the vicinity of 0, the item with (a, b) = (2, 0) would be most appropriate for reducing the uncertainty about ERS_i. Conversely, if a respondent's ERS value was in the neighborhood of –1, an item with (a, b) = (2, –1) would be most useful.

New IRT Model for Measuring ERS

The standard IRT model in Equation 4 addresses one of the limitations of the traditional measure of ERS by clearly separating item (a_k, b_k) and person (ERS_i) effects and allowing items to be differentially useful for measuring ERS (i.e., some items discriminate better between people who are relatively low and those who are relatively high on ERS, and discrimination depends on the item's difficulty). However, the model does not address the remaining three limitations of the traditional ERS measure. Therefore, we extend the standard IRT model and include three novel features in our approach. First, we adapt testlet IRT models, which were originally developed by Bradlow and colleagues (e.g., Bradlow, Wainer, and Wang 1999) for a different purpose, to accommodate substantive correlations among blocks of items that measure the same underlying substantive construct.

Second, we allow for noninvariant ERS properties across groups of respondents, such as different countries, by using a varying item parameter model (i.e., item parameter values for each item are allowed to differ across countries). This provides a unique contribution to multigroup IRT research. To date, all cross-group IRT models have required measurement-invariant anchor items to make the scale of the latent variable common across groups (e.g., Holland and Wainer 1993; May 2006; Reise, Widaman, and Pugh 1993). In other words, the item parameters must be the same in all countries for these anchor items to identify the model. Apart from the difficulties of testing for invariance, there may not be invariant items when many groups are considered. In such cases, existing multigroup IRT models cannot be applied. In our model, there is no longer a need to classify items as invariant or noninvariant.

Finally, we integrate the advanced IRT measurement model with a structural multilevel model, which enables us to study the antecedents of ERS. Previously, researchers have considered only structural multilevel models in connection with the basic IRT model, assuming invariant item parameters across groups (Fox and Glas 2001). We extend the basic measurement model using testlets and varying item parameters and subsequently integrate this model with a structural multilevel model for ERS.

Testlet Structures

Conditional independence is an important assumption in IRT models. It means that for a given respondent, there is no relationship between the respondent's extreme responses

to any pair of Likert items given the latent ERS score. When the ERS measure contains blocks of items that are correlated for substantive reasons (because they measure the same substantive construct), the estimates of the latent ERS score and item parameters will be biased because of the dependence structure between items from the same multi-item scale.

We draw on the educational measurement literature and extend the basic IRT model by incorporating "testlet" effects (Bradlow, Wainer, and Wang 1999). In a series of articles, Bradlow and colleagues have shown the biasing effects of testlet structures on IRT person and item parameters (Bradlow, Wainer, and Wang 1999; Wainer, Bradlow, and Du 2000; Wang, Bradlow, and Wainer 2002). In the current context, each testlet is a multi-item scale, and there are as many testlets as there are multi-item scales. The common content among the items in the multi-item scale is due to the items measuring the same latent marketing construct. Mathematically, the normal ogive model in Equation 4 is adapted as follows:

$$(5)\qquad P(EXTR_{ik} = 1|ERS_i, \psi_{i,r_k}, a_k, b_k) = \Phi[a_k(ERS_i - \psi_{i,r_k} - b_k)],$$

where r_k indicates the testlet of item k. We assume that there are R testlets in total so that $r_k \in \{1, 2, \ldots, r, \ldots, R\}$ and that testlet r contains N_r items. In Equation 5, ψ_{i,r_k} is a person-specific testlet effect, which is independent of ERS_i and the item parameters. It is formulated as a deviation from a person's average ERS value. The parameter ψ_{i,r_k} allows respondents to have a higher ($\psi_{i,r_k} < 0$) or lower ($\psi_{i,r_k} > 0$) probability of giving an extreme response to item k because of the particular testlet r_k (i.e., depending on which substantive construct the item measures). Following Wang, Bradlow, and Wainer (2002), we assume the prior specification $\psi_{i,r_k} \sim N(0, \sigma^2_{\psi_{r_k}})$. That is, we allow for testlet-specific variance parameters.

Cross-Nationally Varying Item Parameters

Measuring ERS in different countries (or different linguistic subgroups) poses additional difficulties, because the item parameters are likely to be noninvariant across countries. The model needs to be able to adjust item characteristics for each item across countries. To accommodate this, we extend recent psychometric models and propose a random-effects analysis of variance structure for the item parameters. Indexing country by j, j = 1, …, J, we use an independent prior specification for a_{kj} and b_{kj}—that is, $a_{kj} \sim N(\tilde{a}_k, \sigma^2_{\tilde{a}})I(a_{kj} \in A)$, $b_{kj} \sim N(\tilde{b}_k, \sigma^2_{\tilde{b}})$—and a multivariate prior for $\tilde{a}_k$ and $\tilde{b}_k$:

$$(6)\quad \xi_k = \begin{bmatrix} \tilde{a}_k \\ \tilde{b}_k \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{\tilde{a}} \\ \mu_{\tilde{b}} \end{bmatrix}, \begin{bmatrix} \sigma^2_{\tilde{a}} & \sigma_{\tilde{a}\tilde{b}} \\ \sigma_{\tilde{a}\tilde{b}} & \sigma^2_{\tilde{b}} \end{bmatrix}\right) = N(\mu_\xi, \Sigma)I(\tilde{a}_k \in A),$$

where A is a bounded interval in $\Re^+$, $I(\cdot)$ is an indicator function, $\log(\mu_{\tilde{a}}) = 0$, $\mu_{\tilde{b}} = 0$, $\Sigma \sim Inv - W(n_0, S)$, $n_0 = 2$, and $S = diag(100, 100)$. In other words, the discrimination and difficulty parameters in a particular country j are draws from independent normal distributions with means of $\tilde{a}_k$ and $\tilde{b}_k$, and the discrimination parameter should be positive. At Level 2, we allow these parameters to be correlated with covariance $\sigma_{\tilde{a}\tilde{b}}$. The prior for the variance–covariance matrix is assumed to be a noninformative inverse-Wishart distribution. We chose this approach for reasons of parsimony. Otherwise, there are too many covariance parameters to be estimated for which we also have to specify priors. Indeed, there is correlation between item parameters within countries, but we did not model this a priori (in the posterior, the parameters can be correlated). A distinction should be made between within-country item correlation and between-country item correlation. The correlation between item parameters across countries has been modeled a priori because at this level, the marginal correlation between the item parameters can be larger. The marginal within-country correlation is small in comparison.

Previously, Janssen and colleagues (2000) considered random-effects specifications for item parameters, though in their article, the grouping was based on items rather than on countries, as in our setting. In addition, they used independent rather than multivariate priors. We combine the random-effects specifications for item parameters with a random-effects structure for ERS (see Fox and Glas 2001).

To summarize, the IRT measurement model for ERS is given by the following:

$$(7)\qquad P(EXTR_{ijk} = 1|ERS_{ij}, \psi_{ij,r_k}, a_{kj}, b_{kj}) = \Phi[a_{kj}(ERS_{ij} - \psi_{ij,r_k} - b_{kj})],$$

$$(8)\qquad a_{kj} \sim N(\tilde{a}_k, \sigma^2_{\tilde{a}})I(a_{kj} \in A),$$

$$(9)\qquad b_{kj} \sim N(\tilde{b}_k, \sigma^2_{\tilde{b}}),$$

$$(10)\qquad [\tilde{a}_k, \tilde{b}_k]^T = \tilde{\xi}_k \sim N(\mu_{\tilde{\xi}}, \Sigma)I(\tilde{a}_k \in A),$$

$$(11)\qquad \psi_{ij,r_k} \sim N(0, \sigma^2_{\psi_{r_k}}),$$

$$(12)\qquad ERS_{ij} \sim N(\beta_{0j}, \sigma^2), \text{ and}$$

$$(13)\qquad \beta_{0j} \sim N(\gamma_{00}, T),$$

where ERS_{ij} denotes the latent ERS score for respondent i in country j (i = 1, …, n_j, j = 1, …, J).

The random-effects specifications for ERS and the item parameters yield an identification problem. Restrictions that fix the mean and variance of the ERS scale in each country are necessary. Each latent ERS country mean can be shifted by changing β_{0j}, as well as by uniformly shifting the country-specific difficulty values b_{kj}. To solve this problem, the latent ERS mean of country j is fixed by restricting the country-specific difficulty parameters in such a way that a common shift of country-specific difficulty values is not possible. This can be done by setting $\Sigma_k b_{kj} = 0\ \forall j$ (Albert 1992). Because this restriction is applied in each country, the mean of the metric of the latent variable is identified through restrictions on the country-specific difficulty parameters. Analogously, the country variances can be shifted by uniform changes in the discrimination parameters. To fix the country variances, we need to impose a restriction such that a common shift of the country-specific discrimination parameters is not possible, which can be done by specifying that, across items, the product of the discrimination parameters equals one in each country j ($\Pi_k a_{kj} = 1\ \forall j$; see Albert 1992). The estimated discrimina-

tion parameters have a product of one because they affect the probability of extreme responding in a multiplicative way, whereas the estimated difficulty parameters sum to zero because their effect is additive (see Equation 4). Thus, the mean and variance of the latent ERS variable in each country are fixed, and the scale remains common because of the simultaneous calibration of the multilevel structures for item parameters and the latent variable. The model allows respondents to be calibrated on the same latent ERS scale even when all items display differential item functioning (DIF) across groups.

The hierarchical Bayesian framework allows for borrowing of strength across countries. That is, the model estimates the distribution of coefficients across the population and combines this information with the responses in a country to derive posterior estimates of country-level parameters (Rossi and Allenby 2003). Previous multigroup IRT research has modeled country means and variances, as well as item parameters, as separate parameters, without borrowing strength across countries. By borrowing strength, we can place less restrictive assumptions on measurement invariance, while retaining the possibility of letting the various parameters fluctuate across countries.

Structural Multilevel Latent Variable Model

Apart from considering the ERS value a bias estimate, survey researchers are also interested in understanding what drives the variation in ERS across people and countries (e.g., Baumgartner and Steenkamp 2001; Greenleaf 1992a, b; Rossi, Gilula, and Allenby 2001). To examine individual and national drivers of ERS, the multilevel model of ERS with testlets can be further extended. We specify the following:

$$\text{ERS}_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \ldots + \beta_{Qj}X_{Qij} + \eta_{ij}, \text{ and} \tag{14}$$

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}W_{1qj} + \ldots + \gamma_{qS}W_{Sqj} + u_{qj}, \tag{15}$$

where X_{1ij} to X_{Qij} are individual-level covariates, W_{1qj} to W_{Sqj} are country-level variables, and η_{ij} and u_j are Level 1 and Level 2 error terms, respectively, with $\eta_{ij} \sim N(0, \sigma^2)$ and $\mathbf{u}_j = (u_{0j}, \ldots, u_{Qj})' \sim N_{Q+1}(0, \mathbf{T})$. Note that the ERS_{ij} term is unobserved and estimated by the IRT model.

Estimation

Combining Equations 7–11, 14, and 15 yields a complex multilevel IRT structure. We used Markov chain Monte Carlo (MCMC) methods to estimate all parameters simultaneously, avoiding the evaluation of high-dimensional integrals (Rossi and Allenby 2003). The MCMC algorithm uses data augmentation to draw samples from the conditional distributions of the parameters (Tanner and Wong 1987). The full conditionals of all parameters can be specified in closed form, and we use a Gibbs sampler to estimate the parameters. Each iteration of the Gibbs sampler consists of sequentially sampling from the full conditional distributions associated with the unknown parameters.[4]

[4]Because of space limitations, we do not present the details of the MCMC scheme. The estimation details appear in the Web Appendix (see http://www.marketingpower.com/jmrfeb08).

How to Correct for ERS Bias

The estimation procedure provides an ERS score, which can be used to adjust survey data for ERS bias. The full conditional distribution of the ERS scores is given by

$$\text{ERS}_{ij}|\mathbf{Z}_{ij}, \mathbf{a}, \mathbf{b}, \psi, \beta_j, \sigma^2 \tag{16}$$

$$\sim N\left[\frac{\sum_{k=1}^{K} a_{kj}(Z_{ijk} + b_{kj} + a_{kj}\psi_{ij,r_k}) + \mathbf{X}_{ij}\beta_j/\sigma^2}{1/\sigma^2 + \sum_{k=1}^{K} a_{kj}^2}, \frac{1}{1/\sigma^2 + \sum_{k=1}^{K} a_{kj}^2}\right],$$

where the parameters Z_{ijk} come from the data augmentation and have a full conditional distribution given by

$$Z_{ijk}|\text{EXTR}_{ijk}, \text{ERS}_{ij}, a_{kj}, b_{kj}, \psi_{ij,r_k} \tag{17}$$

$$\sim \begin{cases} N[a_{kj}(\text{ERS}_{ij} - \psi_{ij,r_k}) - b_{kj}, 1] \\ \text{truncated right by 0 if } \text{EXTR}_{ijk} = 0 \\ N[a_{kj}(\text{ERS}_{ij} - \psi_{ij,r_k}) - b_{kj}, 1] \\ \text{truncated left by 0 if } \text{EXTR}_{ijk} = 1 \end{cases}.$$

Therefore, the ERS scores depend on the testlet parameters, the item parameters, the Level 1 error term, the covariates and associated coefficients, and the augmented data Z_{ijk}. After the ERS scores have been obtained from the MCMC algorithm, adjusting scale scores follows the same logic as the approach that Baumgartner and Steenkamp (2001) and Podsakoff and colleagues (2003) advance. Although different correction procedures can be envisaged, the most straightforward is regressing summated scales or individual items on the ERS score, where the latter option is preferable (Podsakoff et al. 2003). Subsequently, the purified scores can be used in other statistical analyses.

SIMULATION STUDY

Purpose

Before presenting an empirical application of the proposed model and reporting some findings pertaining to the individual and national-cultural drivers of ERS, we perform a simulation study based on synthetic data to compare the performance of our proposed IRT model with the traditional ERS operationalization. We evaluate the traditional model given by Equation 2 and our IRT model with regard to their respective abilities to recover true latent ERS values when there are (1) substantively correlated blocks of items and (2) within- and across-country DIF. In addition, we investigate whether the item parameters of the IRT model can be recovered accurately and whether our IRT model is prone to indicate spurious differences between items and countries when none are present.

Design

We assume that there are 20 countries, with 300 respondents per country. We use 50 items to construct the ERS measure, based on five ten-item "substantive" scales. Thus, there are five testlets. We consider three different testlet

specifications and two specifications about DIF for a total of six different conditions.

We chose respondent-specific testlet parameters to reflect either no, moderately strong, or strong dependencies between the items within a testlet; that is, $\psi_{ij,r} = 0$; $\psi_{ij,r} \sim N(0, .25)$; $\psi_{ij,r} \sim N(0, .5)$, in combination with $ERS_{ij} \sim N(0, 1)$.

For the item parameters, we consider two specifications. As a baseline model, we assume no DIF—that is, identical item parameters across items and countries ($a_{kj} = 1$, $b_{kj} = 0 \forall k, j$). This specification is useful for investigating whether the IRT model might spuriously indicate variation in item parameters across countries when there is no variation. The alternative model allows for DIF—that is, different item parameters within and across countries: $a_{kj} \sim N(\tilde{a}_k, .2^2)$, $b_{kj} \sim N(\tilde{b}_k, .3^2)$, $\tilde{a}_k \sim N(1, .1^2)$, and $\tilde{b}_k \sim N(.5, .1^2)$. These values reflect realistic heterogeneity in item functioning, as we show in our illustration using real data.

Observed binary extreme response patterns **EXTR** are generated from our parameter specifications. We use a root mean square error (RMSE) loss function as our measure of accuracy, in which the deviation between the true and the estimated latent ERS score is squared and summed across individuals. To compare the IRT model and the traditional ERS operationalization, we scale the observed sum score variable so that it has the same mean and variance as the estimated ERS scores from the IRT model. According to Lord (1980, p. 46), the IRT latent score and the true score are "the same thing expressed on different scales of measurement." As a result, we can compare the parameter estimates of the IRT model and the model based on Equation 2.

Results

For estimation of the parameters of interest, we used 30,000 iterations from the Gibbs sampler, after discarding the first 10,000 iterations. The RMSE values for the traditional operationalization of ERS and for the IRT model appear in Table 1.

The message from Table 1 is clear: As the dependencies among the items from the same scale get stronger and DIF increases, the performance of the traditional model deteriorates significantly. In contrast, the latent ERS scores can be recovered much more accurately under the IRT model. As may be expected, RMSE increases somewhat when model complexity increases (i.e., when testlets and DIF are present), but in general, the IRT model performs well and outperforms the traditional model in each condition.

Furthermore, the simulation shows that the item parameters of the IRT model are estimated accurately in every condition (we do not show them because there are no equivalent parameters for the traditional ERS measure). The correlation between the estimated and the true IRT discrimination parameters is .97 ($p < .01$), and the corresponding correlation for the difficulty parameter is .99 ($p < .01$). Finally, there is not much variation in the item parameter estimates across countries when there is no true variation across countries (i.e., $a_{kj} = 1$, and $b_{kj} = 0 \forall kj$). Within- and across-country averages for the discrimination parameters vary between .97 and 1.04, whereas the averages for the difficulty parameters vary between –.04 and .05. Thus, the complex IRT model is not prone to indicate spurious differences between items and countries.

EMPIRICAL APPLICATION

In this section, we present an empirical application to illustrate the IRT model. We estimate the model in a cross-national setting, assess the necessity of allowing for DIF and testlet effects, conduct item parameter validation tests, and investigate individual and cultural drivers of ERS.

The data collection was part of a large multinational study. Two global marketing research agencies, GfK and Taylor Nelson Sofres, collected the data in 26 countries on four continents. The sample in each country was drawn to be broadly representative of the total population in terms of region, age, education, and gender. For countries with high Internet penetration, a Web survey was used. In countries with low Internet penetration, data were collected by mall intercepts in multiple regions/locations. The number of respondents per country varies between 355 (United Kingdom) and 640 (Germany). Given the importance of the United States, the marketing research agencies wanted to have a larger sample for that country (1181 respondents). The total number of respondents is 12,506.

The questionnaire was developed in English and then translated into all local languages by professional agencies, using back translation. To assess ERS, we used a hetero-

Table 1
RECOVERY OF TRUE ERS VALUES

	No Testlet Dependence	*Moderately Strong Testlet Dependence*	*Strong Testlet Dependence*
No DIF	Traditional method: RMSE = 137.4	Traditional method: RMSE = 165.4	Traditional method: RMSE = 182.5
	IRT method: RMSE = 100	IRT method: RMSE = 108.1	IRT method: RMSE = 115.2
DIF	Traditional method: RMSE = 165.9	Traditional method: RMSE = 178.7	Traditional method: RMSE = 191.9
	IRT method: RMSE = 118.9	IRT method: RMSE = 123.2	IRT method: RMSE = 126.1

Notes: The presented RMSE values are relative to the best-performing cell. The IRT method in case of no DIF and no testlet dependence has an RMSE of 21.1, and we normalized this value to 100.

geneous set of 19 multi-item scales and two single items. The total number of items was 100. For all constructs, we used five-point Likert items, and we randomly dispersed the items for each construct throughout the questionnaire. There is a debate in the literature on 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 the desire to look good to others (e.g., evaluation apprehension) or to oneself (e.g., cognitive consistency, ego defense mechanisms). However, randomization may also reduce reliability (Bradlow and Fitzsimons 2001). Information was collected on age (measured in years), gender (1 = women, and 0 = men), and education. In the analyses, we used a within-country median split for education.

RESULTS

Model Selection

On the basis of Equations 7–13, which summarize the full testlet multilevel IRT specification with cross-nationally varying item parameters, we calibrate four nested IRT models. Before conducting these analyses, we estimated a model with a dummy variable V_j in Equation 13; that is, $\beta_{0j} \sim N(\gamma_{00} + \gamma_{01}V_j, T)$, where V_j indicates whether the survey in a given country was a hard copy ($V_j = 1$) or an Internet version ($V_j = 0$). The parameter γ_{01} was not significantly related to ERS in any hierarchical model. The first model (M_1) has cross-nationally invariant item parameters and no testlet structure (i.e., $a_{kj} = a_k$ and $b_{kj} = b_k$, $\forall j$, and $\psi_{ij,r} = 0$, $\forall ij$, r). The second model (M_2) has item parameters that vary across countries and no testlet structure. In other words, Equations 8–10 are specified for the item parameters, but $\psi_{ij,r} = 0$, $\forall ij$, r. The third model (M_3) has cross-nationally invariant item parameters and a testlet structure (i.e., $a_{kj} = a_k$ and $b_{kj} = b_k$, $\forall j$, and Equation 11 for the testlets). Finally, the fourth model (M_4) has both cross-nationally varying item parameters (i.e., Equations 8–10 for the item parameters and Equation 11 for the testlet structure).

To assess which model provides the best fit, we compute the marginal log-likelihood value with importance sampling (Newton and Raftery 1994), log p(**EXTR**), for each ERS measurement model, where **EXTR** contains the binary-coded extreme responses for all items and all respondents. On the basis of the marginal log-likelihood value, we can compute the Bayes factor BF_{xy} for different models M_x and M_y as exp[log p(**EXTR**|M_x) – log p(**EXTR**|M_y)]. Large values for BF_{xy} provide evidence in favor of model M_x. We present the marginal log-likelihoods and the Bayes factors of the model with testlets and varying item parameters versus the other models in Table 2. It is apparent that incorporating the testlet structure and allowing the IRT item parameters to vary across countries both lead to a substantial improvement in model fit.

Item Parameter Variation

The standard ERS measure assumes that each item contributes equally to the overall ERS score, that is, that each item provides an equally good "test" for the ERS value. Both within and across countries, items should function the same. However, this assumption is seriously violated. For illustrative purposes, we plot the average posterior estimated values of the discrimination and difficulty parameters (i.e., a_{kj} and b_{kj}) for all items in China, Germany, and the United States in Figure 2. There is considerable variation in both discrimination and difficulty across items; we also obtain these results for other countries. The notion that items differ in their sensitivity to ERS is a general phenomenon that is not restricted to specific cultures only.

In addition to variation within countries, there is also considerable variation in the item parameters a_{kj} and b_{kj} across countries. For most items, the standard deviation in parameter estimates across countries is approximately .4–.5. Moreover, the cross-national variation in a_{kj} and b_{kj} is not homogeneous across items. To substantiate this, we computed the correlation between the a_{kj} and b_{kj} parameters across countries for each pair of items. Low correlations indicate that a country's standing on a_{kj} or b_{kj} is not consistent across items. The average of the 4950 distinct pairwise correlations for a is –.007 ($p > .1$), whereas the average for b is .011 ($p > .1$).

Item Parameter Stability

We assessed item parameter stability by splitting the within-country samples into equal halves and then estimated the IRT model on both halves. In each split half, the model with testlets and varying item parameters (model M_4) is preferred. The correlation between the two split halves is .78 ($p < .001$) for the discrimination parameters and .87 ($p < .001$) for the difficulty parameters. Thus, both model selection and item parameter estimates are stable across samples, indicating that the differences in difficulty and discrimination across items and countries are replicable and do not simply capture random noise.

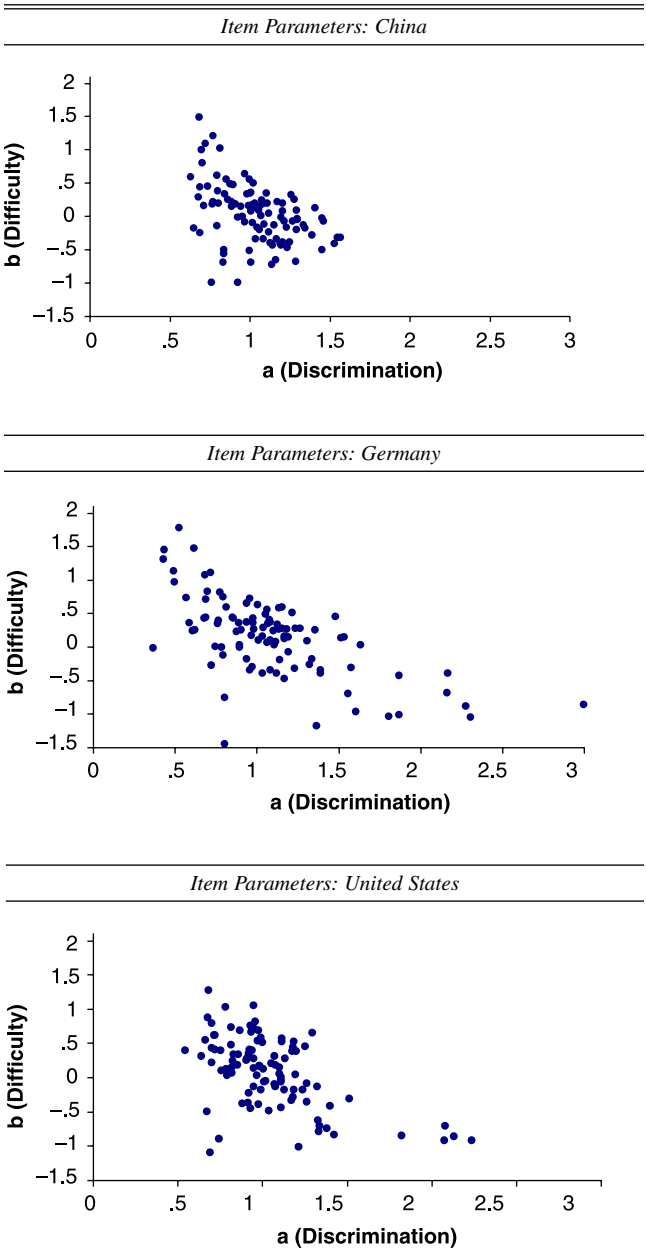
Item Interpretation

In the IRT model, the discrimination parameter a_k and the difficulty parameter b_k have a clear interpretation. The former determines whether an item discriminates well

Table 2
MARGINAL LOG-LIKELIHOOD VALUES AND BAYES FACTORS FOR DIFFERENT IRT MODELS

Model	*Marginal Log-Likelihood*	*Bayes Factor*
M_1: Invariant item parameters and no testlets	–461,384	BF_{41} = Exp(40,859)
M_2: Varying item parameters and no testlets	–449,088	BF_{42} = Exp(28,563)
M_3: Invariant item parameters and testlets	–432,861	BF_{43} = Exp(12,336)
M_4: Varying item parameters and testlets	–420,525	—

Figure 2
ITEM PARAMETERS FOR CHINA, GERMANY, AND THE UNITED STATES



between people who are relatively high and those who are relatively low on ERS, whereas the latter refers to the probability that an item elicits an extreme response. However, are there characteristics of items that are systematically related to a_k and b_k? To examine this, we correlated a_k and b_k within countries with several item characteristics. Given the relative absence of theory to guide us, we approach this issue in an exploratory manner. Using the strength of the data set, we then performed a meta-analysis on the correlations across our sample countries using the method of adding Zs (Rosenthal 1991). We consider the number of words in an item (Wang, Bradlow, and Wainer 2004); the number of characters in the item (we excluded the Asian languages because sentence structures are different); the way the item is worded, either positively or negatively (Wong, Rindfleisch, and Burroughs 2003); and an item's deviation from the midpoint of the scale (Baumgartner and Steenkamp 2001).

We find that the difficulty parameter is negatively correlated with the item's absolute deviation from the midpoint ($r = –.730$, $p < .001$). This result has face validity because when the absolute deviation from the scale midpoint is large, the item has elicited many extreme responses so that the difficulty parameter should be negative. More words ($r = –.276$, $p < .001$) and more characters ($r = –.295$, $p < .001$) are also negatively associated with the difficulty parameter.

We further find that items that discriminate better between people who are relatively high and those who are relatively low on ERS deviate more strongly from the midpoint of the scale ($r = .180$, $p < .001$), are longer ($r = .436$, $p < .001$), contain more characters ($r = .441$, $p < .001$), and are worded positively ($r = –.117$, $p < .001$).

DRIVERS OF ERS

Sociodemographic Variables

Previous research has investigated whether extreme responding is related to characteristics of individuals. The sociodemographic variables age, gender, and education have attracted the most attention (Greenleaf 1992a, b; Marín, Gamba, and Marín 1992). We also include these three sociodemographics in our model. The results of prior research have not been very consistent, and we use the improved measurement of ERS and the large multinational data set to investigate whether the sociodemographic variables studied are reliably related to ERS.

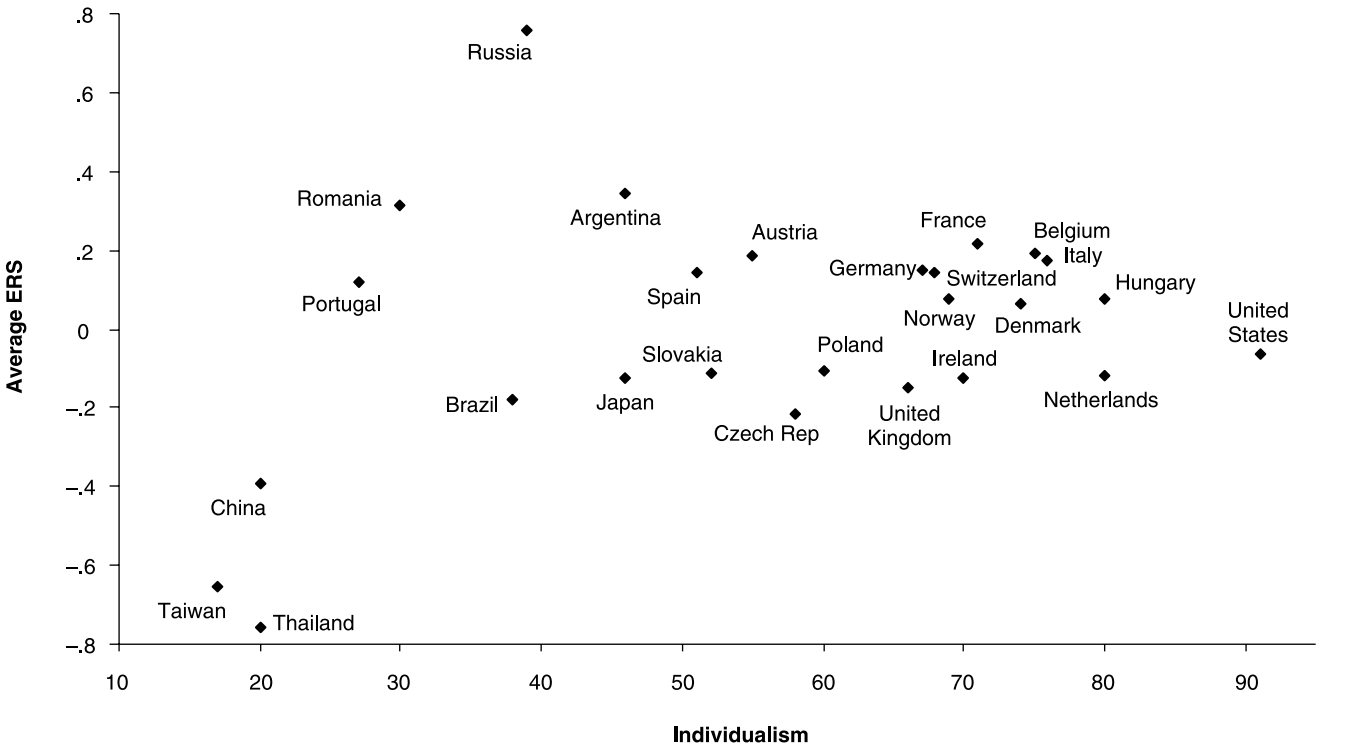
Cultural Drivers of ERS

Extreme response style may differ systematically not only between individuals but also between countries (Baumgartner and Steenkamp 2001; Chen, Lee, and Stevenson 1995; Grimm and Church 1999; Johnson et al. 2005). Figure 3 displays the average ERS value in each country as a deviation from the grand ERS mean.

What gives rise to these cross-national differences? We propose that a country's culture is a major driver of country differences in ERS. To investigate the effect of culture on ERS, we employ Hofstede's (2001) framework of cultural dimensions and the classification of countries on those dimensions. For example, Figure 3 shows that countries low on individualism have relatively low ERS scores. Subsequently, we develop our hypotheses more formally.

Desire for uniqueness and independence are core elements of cultural individualism (Oyserman, Coon, and Kemmelmeier 2002). In individualist societies, a person's attitudes are regulated largely by individual preferences, and the expression of unique opinions is valued (Chen, Lee, and Stevenson 1995). An individual's identity is clearly distinct from that of other people (Hofstede 2001). In contrast, in collectivist societies, attitudes are relatively more heavily influenced by society's preferences. These cultures are characterized by an interdependent self-concept and encourage modesty and harmony (Triandis 1989). Therefore, we expect a positive relationship between a country's degree of individualism and ERS.

Figure 3
ERS VALUES AND INDIVIDUALISM SCORES



At the individual level, studies in psychology have repeatedly shown that extreme responding is positively related to intolerance of ambiguity, rigidity, and need for certainty (for a review, see Baumgartner and Steenkamp 2001). Hofstede (2001) argues that differences in intolerance of ambiguity are also a cultural characteristic (termed "uncertainty avoidance"). Uncertainty avoidance measures the degree to which societies are made nervous and feel threatened by uncertain, risky, ambiguous, or undefined situations. To avoid such situations, they tend to adopt rigid attitudes and rules. Thus, we expect that there is a positive relationship between uncertainty avoidance and ERS.

Cultural masculinity/femininity is defined as the degree to which a society is characterized by assertiveness versus nurturance (Hofstede 2001). Masculine societies place great emphasis on achievement and ambition and encourage assertiveness and decisive/daring behavior, which should lead to a tendency to select the strongest available choices on Likert rating scales. Feminine societies value social harmony, gentleness, and modesty, which implies that ERS should be less common.

For completeness, we also include power distance in the model, though we do not have strong a priori expectations about its effect on ERS. Johnson and colleagues (2005) theorize that it is positively related to ERS because high-power-distance societies demand decisiveness and definiteness in communications by superiors, whereas subordinates should respond modestly, if not deferentially. However, this implies that the effect depends on a person's position in the social hierarchy, and thus, in general, a null effect seems more likely.

In summary, we hypothesize that ERS is greater in countries whose culture is characterized by higher levels of individualism, uncertainty avoidance, and masculinity. We predict no relationship for power distance.

Results

In the simultaneous estimation of the measurement (Equations 7 to 11) and structural model (Equations 14 and 15), we first considered the necessity of including random coefficients for the Level 1 predictors. Raudenbush and Bryk (2002) recommend constraining slope coefficients that do not display random variation across countries to be fixed for increased parameter stability and efficiency. We found significant variation across countries for gender and education but not for age. Thus, we constrained the coefficients for age to be fixed, and we specified the slopes for the other two variables as random. The results for this model appear in Table 3.

The sociodemographic variables explain approximately 2% of the Level 1 variance. Women tend to score higher on ERS than men ($\gamma_{03} = .0324$), and both younger and older people are more prone to respond extremely ($\gamma_{01} = -.1463$, $\gamma_{02} = .1278$). For education, we found no cross-nationally generalizable effect, though there was significant random variation across countries.

Culture plays an important role in explaining cross-national differences in ERS. The four culture dimensions

Table 3
DRIVERS OF ERS

	Coefficient	*SD*
γ_{00} (constant)	–1.2411[a]	.0554
Sociodemographic Variables		
γ_{01}(age)	–.1463[a]	.0469
γ_{02} (age × age)	.1278[a]	.0411
γ_{03} (gender [1 = female; 0 = male])	.0324[a]	.0093
γ_{04} (education)	.0010	.0103
National-Cultural Variables		
γ_{05} (individualism)	.0037[a]	.0021
γ_{06} (uncertainty avoidance)	.0052[a]	.0020
γ_{07} (masculinity)	.0030[a]	.0017
γ_{08} (power distance)	–.0024	.0558
Variance Parameter		
σ^2 (Level 1 variance)	.5661[a]	.0084

[a]Indicates that the 95% posterior probability interval excludes zero.

explained 59% of the between-country variance in ERS. As we hypothesized, ERS is positively related to national-cultural individualism (γ_{05} = .0037), uncertainty avoidance (γ_{06} = .0051), and masculinity (γ_{07} = .0029). At first, the difference in direction between masculinity and gender may seem counterintuitive. However, it is well known that relationships between the same variables may be different at the individual and cultural levels because the underlying mechanisms are different (Hofstede 2001, p. 216). Moreover, masculine and feminine do not refer in any simple way to fundamental traits of personality but rather to the historical gender role patterns in a society. It would be inappropriate to equate this with contemporary gender roles. As we expected, ERS is not related to power distance. Johnson and colleagues (2005) also recently investigated the relationship between Hofstede's dimensions and extreme responding across 19 countries. They found no statistically significant effects in their initial analysis using the original Hofstede scores, though they found positive effects for power distance and masculinity when they trichotomized the scores of the countries in their sample. This suggests that the proposed methodology can help reveal drivers of ERS that cannot be observed with other ERS measures.

CONCLUSIONS

In the introduction, we identified several contributions of this research to the study of ERS. We structure our conclusions around these contributions. First, our new, IRT-based method improves on the traditional ERS method by allowing different items to be differentially useful for measuring ERS and by accommodating the possibility that an item's usefulness differs across groups (e.g., countries). Our simulations show that ignoring DIF within and across countries leads to seriously biased results, and our large-scale empirical study provides strong evidence that survey items do not provide equally useful information about ERS, either within or across countries. People differ in their tendency to use the extremes of the rating scale, and items also differ in the extent to which they elicit extreme responses, both nationally and cross-nationally. The cross-classified character of IRT—that is, disjunct item and person parameters—is well suited to capture this interactive phenomenon, whereas the use of simple sum scores (Equation 2) is rendered problematic by our findings. The results provide support for the notion that stylistic responding is best understood as an interaction of personal dispositions and item characteristics (Podsakoff et al. 2003). A unique feature of our model is that each item is allowed to function differently across countries. Thus, measurement invariance for item parameters is relaxed.

Second, unlike the traditional model, our model allows researchers to purge item scores of ERS even when they only have correlated items measuring substantive constructs. In a simulation study, we showed that ignoring the correlation between items biases the traditional ERS estimate, whereas the inclusion of testlets effectively controls for this problem.

Third, our model integrates the advanced IRT measurement model with a structural hierarchical model for studying the antecedents of ERS. Applying this integrated IRT–hierarchical model to a large data set involving 12,500 consumers from 26 countries, we find that the sociodemographic variables studied have a minor influence on ERS, but culture exerts a strong and predictable effect on ERS. Implications for international marketing are evident because ERS differences might bias comparisons between countries.

There are several promising avenues for further research. First, most research on individual-level drivers of ERS has examined sociodemographics. The results are often inconsistent and the effect sizes small. Further research could examine more fundamental characteristics of individuals, such as personality factors or value priorities. Second, our model assumes that the items in each survey are the same. However, it often happens that there are both common and country-specific items (e.g., May 2006). Our model could be extended to accommodate such situations. Another option is to include "No Answer" options in the survey. It may sometimes be more valid to allow such answers rather than forcing respondents to provide an answer on the rating scale. Further research could integrate our ERS model with response models that have been developed for such situations (see Bradlow and Zaslavsky 1999). Finally, the highest-information measurement method for ordinal data is the graded-response (ordinal) IRT model. However, for such a model to identify ERS and ERS properties of items accurately, it should include all relevant response styles. Further research could work on the specification and interpretation of such generalized response models. Although many issues require additional research, we hope that this article stimulates marketing researchers to pay more careful attention to the issue of ERS in both domestic and international survey research.

REFERENCES

Albert, James H. (1992), "Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling," *Journal of Educational Statistics*, 17 (Fall), 251–69.

Ansari, Asim, Kamel Jedidi, and Sharan Jagpal (2000), "A Hierarchical Bayesian Methodology for Treating Heterogeneity in Structural Equation Models," *Marketing Science*, 19 (Fall), 328–47.

Bachman, Jerald G. and Patrick M. O'Mally (1984), "Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles," *Public Opinion Quarterly*, 48 (Summer), 491–509.

Baumgartner, Hans and Jan-Benedict E.M. Steenkamp (2001), "Response Styles in Marketing Research: A Cross-National Investigation," *Journal of Marketing Research*, 38 (May), 143–56.

——— and ——— (2006), "An Extended Paradigm for Measurement Analysis of Marketing Constructs Applicable to Panel Data," *Journal of Marketing Research*, 43 (August), 431–42.

Bearden, William O. and Richard G. Netemeyer (1999), *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*, 2d ed. Newbury Park, CA: Sage Publications.

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.

———, Howard Wainer, and Xiaohui Wang (1999), "A Bayesian Random Effects Model for Testlets," *Psychometrika*, 64 (2), 153–68.

——— and Alan M. Zaslavsky (1999), "A Hierarchical Latent Variable Model for Ordinal Data from a Customer Satisfaction Survey with 'No Answer' Responses," *Journal of the American Statistical Association*, 94 (March), 43–52.

Chen, Chuansheng, Shin-ying Lee, and Harold W. Stevenson (1995), "Response Style and Cross-Cultural Comparisons of Rating Scales Among East Asian and North American Students," *Psychological Science*, 6 (May), 170–75.

Fox, Jean-Paul and Cees A.W. Glas (2001), "Bayesian Estimation of a Multilevel IRT Model Using Gibbs Sampling," *Psychometrika*, 66 (2), 269–86.

Greenleaf, Eric A. (1992a), "Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles," *Journal of Marketing Research*, 29 (May), 176–88.

——— (1992b), "Measuring Extreme Response Style," *Public Opinion Quarterly*, 56 (Fall), 328–51.

Grimm, Stephanie D. and A. Timothy Church (1999), "A Cross-Cultural Study of Response Biases in Personality Measures," *Journal of Research in Personality*, 33 (4), 415–41.

Hambleton, Ronald K. and Hariharan Swaminathan (1985), *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.

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

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

Hui, Harry C. and Harry C. Triandis (1989), "Effects of Culture and Response Format on Extreme Response Style," *Journal of Cross-Cultural Psychology*, 20 (September), 296–309.

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.

Johnson, Timothy R. (2003), "On the Use of Heterogeneous Thresholds Ordinal Regression Models to Account for Individual Differences in Extreme Response Style," *Psychometrika*, 68 (4), 563–83.

———, Patrick Kulesa, Young Ik Cho, and Sharon Shavitt (2005), "The Relation Between Culture and Response Styles," *Journal of Cross-Cultural Psychology*, 36 (March), 264–77.

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

Lord, Frederick M. (1980), *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Marín, Gerardo, Raymond J. Gamba, and Barbara V. Marín (1992), "Extreme Response Style and Acquiescence Among Hispanics," *Journal of Cross-Cultural Psychology*, 23 (December), 498–509.

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.

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

Oyserman, Daphna, Heather M. Coon, and Markus Kemmelmeier (2002), "Rethinking Individualism and Collectivism: Evaluation of Theoretical Assumptions and Meta-Analyses," *Psychological Bulletin*, 128 (January), 3–73.

Paulhus, Delroy L. (1991), "Measurement and Control of Response Bias," in *Measures of Personality and Social Psychological Attitudes*, John P. Robinson, Phillip R. Shaver, and Lawrence S. Wright, eds. San Diego: Academic Press, 17–59.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff (2003), "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology*, 88 (October), 879–903.

Raudenbush, Stephen W. and Anthony S. Bryk (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

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.

Rosenthal, Robert (1991), *Meta-Analytic Procedures for Social Research*. Newbury Park, CA: Sage Publications.

Rossi, Peter E. and Greg M. Allenby (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22 (3), 304–328.

———, 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.

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

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.

Triandis, Harry C. (1989), "The Self and Social Behavior in Differing Cultural Contexts," *Psychological Review*, 96 (July), 506–520.

Van Herk, Hester, Ype H. Poortinga, and Theo M.M. Verhallen (2004), "Response Styles in Rating Scales: Evidence of Method Bias in Data from Six EU Countries," *Journal of Cross-Cultural Psychology*, 35 (May), 346–60.

Wainer, Howard, Eric T. Bradlow, and Z. Du (2000), "Testlet Response Theory: An Analog for the 3-PL Useful in Testlet-Based Adaptive Testing," in *Computerized Adaptive Testing, Theory and Practice*, W.J. van der Linden and C.A.W. Glas, eds. Boston: Kluwer-Nijhof, 245–70.

Wang, Xiaohui, Eric T. Bradlow, and Howard Wainer (2002), "A General Bayesian Model for Testlets: Theory and Applications," *Applied Psychological Measurement*, 26 (March), 109–128.

———, ———, and ——— (2004), "User's Guide for SCORIGHT (Version 3.0): A Computer Program for Scoring Tests Built of Testlets Including a Module for Covariate Analysis," Educational Testing Service Technical Report RR-04-49, Princeton, NJ.

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.