Journal of Personality and Social Psychology
2012, Vol. 103, No. 3, 543–564

0022-3514/12/$12.00 DOI: 10.1037/a0029394

Analysis of Sensitive Questions Across Cultures: An Application of Multigroup Item Randomized Response Theory to Sexual Attitudes and Behavior

Martijn G. de Jong
Erasmus University Rotterdam

Rik Pieters
Tilburg University

Stefan Stremersch
Erasmus University Rotterdam and Universidad de Navarra

Answers to sensitive questions are prone to social desirability bias. If not properly addressed, the validity of the research can be suspect. This article presents multigroup item randomized response theory (MIRRT) to measure self-reported sensitive topics across cultures. The method was specifically developed to reduce social desirability bias by making an a priori change in the design of the survey. The change involves the use of a randomization device (e.g., a die) that preserves participants' privacy at the item level. In cases where multiple items measure a higher level theoretical construct, the researcher could still make inferences at the individual level. The method can correct for under- and overreporting, even if both occur in a sample of individuals or across nations. We present and illustrate MIRRT in a nontechnical manner, provide WinBugs software code so that researchers can directly implement it, and present 2 cross-national studies in which it was applied. The first study compared nonstudent samples from 2 countries (total $n = 927$) on permissive sexual attitudes and risky sexual behavior and related these to individual-level characteristics such as the Big Five personality traits. The second study compared nonstudent samples from 17 countries (total $n = 6{,}195$) on risky sexual behavior and related these to individual-level characteristics, such as gender and age, and to country-level characteristics, such as sex ratio.

Keywords: social desirability, sensitive questions, randomized response, cross-cultural survey, mating theories

The private nature of sensitive topics such as drug use, alcohol consumption, and sexual behavior makes it challenging to collect objective, archival information on these important behaviors. Some sensitive behaviors may be assessed by means of biomarkers. However, such methods are usually intrusive and expensive, and they cannot be applied to large and geographically dispersed populations (Catania, Gibson, Chitwood, & Coates, 1990). Thus, self-reports are often the method of choice for psychologists who study individual differences in socially sensitive behaviors.

Self-reports of sensitive questions suffer from a number of well-known methodological limitations (Catania et al., 1990; Fenton, Johnson, McManus, & Erens, 2001; Tourangeau & Yan, 2007). The most important limitation is social desirability bias in the self-report measures due to the sensitive nature of the questions involved (Weinhardt, Forsyth, Carey, Jaworski, & Durant, 1998). Embarrassment, cultural taboos, fear of reprisals, and even bragging can lead to discrepancies between the actual and reported behavior. The sensitivity of self-reports may also vary across sociodemographic characteristics, such as gender and age (David & Knight, 2008; Laumann, Paik, & Rosen, 1999) and cultural groups (Lalwani, Shavitt, & Johnson, 2006; Lalwani, Shrum, & Chiu, 2009). This is especially the case when full anonymity of participants' answers to the questions cannot be guaranteed. As a result, the validity of most studies on sensitive topics that use self-report measures is unknown, and minimizing measurement error is a key challenge (Catania et al., 1990; Fenton et al., 2001). In a review and meta-analysis of the literature on sensitive questions in surveys, Tourangeau and Yan (2007, p. 878) stressed that "the need for methods of data collection that elicit accurate information is more urgent than ever." The present research follows up on this call.

This article was published Online First July 23, 2012.

Martijn G. de Jong, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, the Netherlands; Rik Pieters, Department of Marketing, Tilburg School of Economics and Management, Tilburg University, Tilburg, the Netherlands; Stefan Stremersch, Erasmus School of Economics, Erasmus University Rotterdam, and IESE Business School, Universidad de Navarra, Barcelona, Spain.

This research was supported by the Netherlands Organization for Scientific Research (Grant 016.105.116 titled "State-of-the-art cross-cultural research methods" to Martijn G. de Jong), the Erasmus Healthcare Business Center, the Erasmus Research Institute for Management, and the Marketing Science Institute. We thank Nuno Camacho, Sheila M. Jose, Bram van den Berg, and Jan Benedict E. M. Steenkamp for helpful comments.

Correspondence concerning this article should be addressed to Martijn G. de Jong, Erasmus School of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, the Netherlands. E-mail: mgdejong@ese.eur.nl

Various methodologies have been developed to either prevent or correct social desirability bias in self-reports of sensitive topics. A common approach to prevent social desirability bias is to guarantee anonymity to the participants in the introduction of the questionnaire, which may work relatively well for single-shot surveys, student samples, and moderately sensitive questions (see for an example, Schmitt, 2005). However, it is hard to establish similar levels of anonymity and trust in other cases, such as when longitudinal or panel surveys are conducted, adult panel members are surveyed, and the topics are very sensitive. In longitudinal and panel research, already sensitive topics may become more sensitive because the answers of individuals across multiple surveys are linked and participants know this. de Jong, Pieters, and Fox (2010) found that anonymity guarantees did not help in a nationally representative survey study on the sexual desires for commercial sex among adult participants. These desires were underreported when direct questions were used, despite strict guarantees of anonymity. Moreover, the level of underreporting varied as a function of sociodemographic variables, including age and gender. Older participants and women underreported more than men and younger participants.

The bogus pipeline technique (Roese & Jamieson, 1993) that aims to prevent socially desirable responding from tainting the responses to survey questions is problematic as well. In a bogus pipeline study, participants are connected to a fake lie detector and, to encourage truthfulness, are told that it can sense dishonesty. Although this method has been used for small-scale student samples under controlled conditions, it cannot be applied in regular surveys because it is expensive to implement, particularly in cross-cultural research. Moreover, the technique deceives participants, thus violating the codes of conduct of professional survey organizations and market research companies that collect the data.

Besides techniques that try to prevent social desirability bias, a popular approach is to include a social desirable responding (SDR) or lie scale in the survey to correct for bias in the subsequent statistical analyses after the data have been collected (Paulhus, 2002). However, there is evidence that attempts at correcting for social desirability with SDR scales reduce validity because SDR scales measure not only response style but also substantive traits and states, such as tendencies to yield to social pressure (Ellingson, Smith, & Sackett, 2001; Ones, Viswesvaran, & Reis, 1996; Piedmont, McCrae, Riemann, & Angleitner, 2000; D. B. Smith & Ellingson, 2002; Tourangeau & Yan, 2007). Thus, attempts to correct for social desirability bias inadvertently introduce other forms of bias. Moreover, because the method assumes that the direction of the bias is the same for all participants, such postsurvey corrections cannot be used if some participants overreport (brag) and others underreport. Catania et al. (1990) speculated that tendencies to over- and underreport may vary across cultures because of varying social norms. If the claim is accurate, then it threatens the validity of the method even more.

To address shortcomings in prior methods, psychometricians recently introduced a new procedure for measuring individual differences in sensitive behaviors and attitudes: item randomized response theory (IRRT; Böckenholt & van der Heijden, 2007; de Jong et al., 2010; Fox, 2005). The procedure has two components. First, it involves an a priori change in the survey design that guarantees complete anonymity to participants at the level of their true answers to the specific items in the questionnaire. This is accomplished by providing participants with a randomization device for answering. The guaranteed anonymity encourages participants to answer sensitive questions truthfully. Second, it uses a psychometric model to extract participants' true scores on the higher order construct that underlies the specific items. Hence, IRRT integrates methods of randomized response (RR) data collection (Fox & Tracy, 1986; Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005; Warner, 1965) with item response theory (IRT) for data analysis (Fraley, Waller, & Brennan, 2000). De Jong et al. show that in terms of improved nomological validity and bias reduction, this generates better results than post hoc corrections with SDR scales. The method can reduce both under- and overreporting and can deal with situations where some participants underreport and others overreport. Intuitively, these features make IRRT better suited to reduce social desirability bias in the study of sensitive behavior across cultures.

This article explores IRRT in a cross-national setting. It reports the results of two studies. The first study examined permissive sexual attitudes and risky sexual behavior for samples of adult participants from two countries, Spain and the Netherlands. This study enabled inferences about person-level correlates of the permissive sexual attitudes and risky sexual behaviors, in particular sociodemographics and personality. The second study concerned permissive sexual behaviors for samples of adult participants from 17 countries, namely, Belgium, Brazil, Canada, Denmark, Estonia, France, Germany, India, Italy, Japan, Netherlands, Poland, Portugal, Singapore, Switzerland, United Kingdom (UK), and the United States (US). The larger sample of countries in Study 2 allowed inferences about country-level correlates of the behaviors, in addition to the person-level correlates. To the best of our knowledge, this is the first research that applied IRRT to cross-national survey data. Henceforth, the methodology is labeled as MIRRT (multigroup IRRT).

Whereas most of the psychometric literature on the topic is very technical, this article attempts to present MIRRT in an accessible and nonmathematical manner. The focus is primarily on the "how" of the approach. According to a recent editorial article in this journal (Simpson, 2009, p. 60), "There have been many significant advances in research methods and techniques during the past decade, and social and personality researchers are uniquely situated among social and behavioral scientists to take full advantage of these new methodological advances." Thus, in the spirit of Fraley et al. (2000), the present research aims, by providing new analytic strategies, to contribute to the methodological toolbox of social and personality psychologists who study sensitive topics across cultures. To enable wider usage of the proposed methodology, Appendix A provides software code in the free program WinBugs (Lunn, Thomas, Best, & Spiegelhalter, 2000). In addition, the current article aims to advance the field of sex- and health-related research that has often been constrained by limitations to measurement validity. To appreciate the advantages of the proposed MIRRT method, we list here its key benefits. The MIRRT method

1. Protects anonymity of participants at the item level.

2. Does not deceive participants.

3. Can control for both over- and underreporting at the same time.

4. Can identify participants who do not adhere to the procedure.

5. Can analyze multiple samples (countries, cultures) simultaneously.

6. Can link sensitive attitudes and behavior to individual-level characteristics of the participants (thus allowing a mapping of the antecedents and consequences of the sensitive behavior).

There are several reasons why this research focused on sexual attitudes and behavior. First, self-reported sexual attitudes and behavior are frequently used to understand fundamental human mating strategies (Buss & Schmitt, 1993; Eagly & Wood, 1999; Li & Kenrick, 2006). For instance, research has focused on individual differences in restricted versus unrestricted mating orientations (Simpson & Gangestad, 1991, 1992), gender differences in sexual behavior (Buss & Schmitt; R. D. Clark & Hatfield, 1989), mate selection (Kenrick, Keefe, Bryan, Barr, & Brown, 1995), and the relationship between culture and sociosexuality and desire for variety (Schmitt, 2003, 2005). In addition, self-report is the method of choice in research on sexual risk taking (Cooper, 2010). Second, societies typically develop social norms on sexual attitudes and behavior, making socially desirable responding in self-reports about them likely. The social norms also often vary across societies. Finally, to prevent and manage life-threatening sexually transmitted diseases, public policy makers also take a key interest in mapping the antecedents of sexual attitudes and behavior. Sexually transmitted diseases are the prime preventable causes of infertility and sexual dysfunctions, and they are associated with chronic health problems and untimely death. In addition, avoiding risky sexual behavior also reduces social problems associated with divorce, teenage childbirth, unplanned pregnancy, and prostitution. This study examined sexual attitudes and behavior that may contribute to such problems across countries and adult participants.

IRRT

IRRT integrates data collection and data analysis of sensitive questions. During data collection, participants use a randomization device, like a coin or die, before answering a specific sensitive question. The outcome of the randomization device determines whether participants should answer the question truthfully or provide a forced answer that is given by a subsequent outcome of the randomization device. This produces an item randomized response (IRR). IRRT has gone through several stages in its development, from the initial model (M_1) to the currently proposed model (M_{10}). Table 1 illustrates these.

In view of the empirical application, let us take as a running example, the single item “During the last six months, have you had sex with a sex worker?” Initially, assume a binary “yes or ‘no” response format (M_1 in table 1). When a coin is used for randomization, participants flip it before answering the question. The researcher does not observe the flip, and participants only have to answer the item truthfully if the coin comes up, say, heads. However, if the coin comes up tails, participants always answer *yes* irrespective of whether that answer is true or not. This protects the participant because the researcher has not seen the flip and therefore cannot know if *yes* means that the participant has engaged in the behavior or if the coin came up tails.

Although the researcher cannot know the true behavior of a *specific participant*, the model M_1 can calculate the *proportion of the sample* that has engaged in the specific behavior. Thus, this model provides aggregate (sample-level) results but not disaggregate (individual-level) results (see column 5 in Table 1). That is,

Table 1
Evolution of Item-Randomized Response (IRRT) Models

	Characteristics of data collection and analysis					
	Data collection		Data analysis			
Model	No. questions	Response scale	Account for procedure non adherence	Analysis level	Multigroup	Equation
M_1	Single	Binary	No	Sample	No	$P(Y = 1) = p_1\mu + (1 - p_1)p_2$
M_2	Single	Polytomous	No	Sample	No	$P(Y = c) = p_1\mu_c + (1 - p_1)p_{2,c}$
M_3	Single	Binary	Yes	Sample	No	M_1 + nonadherence detection
M_4	Single	Polytomous	Yes	Sample	No	M_2 + nonadherence detection
M_5	Multiple	Binary	No	Individual	No	$P(Y_{ik} = 1) = p_1\pi_{ik} + (1 - p_1)p_2$ $\pi_{ik} = 1/[1 + \exp(-\alpha_k(\theta_i - \beta_k)]$
M_6	Multiple	Polytomous	No	Individual	No	$P(Y_{ik} = c) = p_1\pi_{ikc} + (1 - p_1)p_{2,c}$ $\pi_{ikc} = 1/[1 + \exp(-\alpha_k(\theta_i - \beta_{k,c-1})] - 1/[1 + \exp(-\alpha_k(\theta_i - \beta_{k,c})]$
M_7	Multiple	Binary	Yes	Individual	No	M_5 + latent class for nonadhering participants
M_8	Multiple	Polytomous	Yes	Individual	No	M_6 + latent class for nonadhering participants
M_9	Multiple	Polytomous	No	Individual	Yes	$P(Y_{ijk} = c) = p_1\pi_{ijkc} + (1 - p_1)p_{2,c}$ $\pi_{ijkc} = 1/[1 + \exp(-\alpha_k(\theta_{ij} - \beta_{k,c-1})] - 1/[1 + \exp(-\alpha_k(\theta_{ij} - \beta_{k,c})]$
M_{10}	Multiple	Polytomous	Yes	Individual	Yes	M_9 + latent class for nonadhering participants

Note. Y is the observed score on the question; p_1 is the probability of having to answer honestly, μ_c is the true sample probability of response c when directly and honestly answering the question; $p_{2,c}$ is the probability of some forced response c given that a forced response has to be given; θ_i is individual i's latent trait; θ_{ij} is the latent trait of individual i in country j; α_k is the discrimination parameter; β_{kc} is the difficulty parameter of category c.

participants who have engaged in the behavior will have answered *yes* regardless of the outcome of the randomization device, and half of the participants who have *not* engaged in the behavior will have also answered *yes*. The formula to calculate the true proportion that has performed the behavior from the observed proportion is found in the first row of the last column of Table 1. In this formula, p_1 is the probability of having to give a truthful answer. If a regular coin is used, $p_1 = .50$ (50% chance). Furthermore, p_2 is the probability of a forced response of *yes* in case a forced response needs to be given. In the current example with the coin, $p_2 = 1$ (100% *yes* responses in case of tails). Thus, if 60% of the sample responded *yes* to the question, the actual proportion that has had sex with a sex worker in the last six months is 20% because $0.6 = (0.5 \times \text{actual proportion}) + (0.5 \times 1)$.

Model M_1 is the basic randomized response model. It is useful but has limitations in theory testing and public policy implementation. Therefore, four improvements have been developed: (a) polytomous responses, (b) nonadherence to the procedure, (c) individual-level analysis, and (d) multigroup comparisons.

Polytomous Responses

Social scientists often desire more fine-grained, graded response options rather than binary (yes/no) answers. The question can be, for instance: "During the last six months, how *often* have you had sex with a sex worker?" with multiple response options ranging from *Never* to *Often*. To accommodate this, a polytomous version of the randomized response model was developed earlier (model M_2; Abul-Ela, Greenberg, & Horvitz, 1967). Note that the randomization device in this case is no longer a coin but is either a spinner or a die. Participants turn a spinner or cast a die and provide either their true answer to the question or one of the forced responses, depending on the number that appears. The empirical applications in the current project used a die (see Figure 1). With it, $p_1 = 4/6$, $p_{2,1} = p_{2,2} = p_{2,3} = p_{2,4} = 1/6$, and $p_{2,5} = 2/6$. Note that unequal probabilities to respond (e.g., here 1/6 vs. 2/6) can be easily handled as long as they are known. The formulas to infer the true responses from the observed responses to the question change slightly but not much (see the second row of the last column in Table 1). Again, p_1 is the probability that a truthful answer needs to be given. If a forced response needs to be given (with probability $1 - p_1$), category c has probability $p_{2,c}$ to be selected. The third column of Table 1 indicates this improvement to handle polytomous questions.

Nonadherence to the Procedure

Even though their anonymity is completely protected by the randomized response procedure, some participants may still not adhere to the instruction. Some people may consistently report the least incriminating response categories, whatever the conditions are. Several studies have qualitatively documented this phenomenon (S. J. Clark & Desharnais, 1998). Edgell, Himmelfarb, and Duncan (1982) showed that when probed about homosexual responses, 25% of the participants who had to answer *yes* according to the RR design still gave a *no* response. Boeije and Lensvelt-Mulders (2002) observed that most participants in their study found it difficult to give a false *yes* response that the procedure called for and instead opted for the *no* response. Most often, such nonadherence to the procedure occurs because the question is so sensitive that the participant wants to make sure she or he is not incriminated and thus answers *no* although the randomization device calls for a *yes*. Nonadherence reduces the validity of randomized response methods. In one approach to assess and control for it, researchers have developed models M_3 and M_4 (S. J. Clark & Desharnais, 1998; Moshagen, Musch, Ostapczuk, & Zhao, 2010) that rely on manipulation of p_1 and p_2 across experimental groups. The idea is that when the probability of having to answer *yes* due to the randomization device is high, it is easier to truthfully respond *yes* as well. From the differences between experimental groups in inferred true responses, the proportion of the sample not adhering to the procedure can be derived. In this research, we relied on a different approach described later.

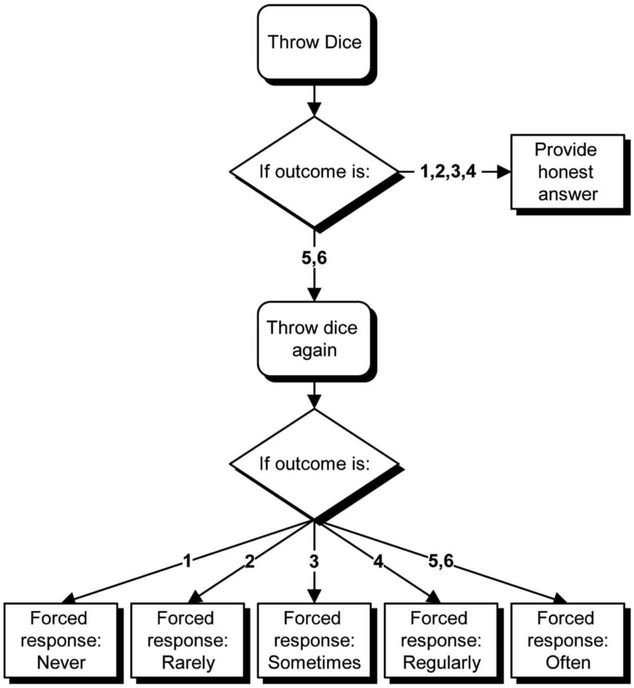

Figure 1. Flow of randomized response data collection.

Individual-Level Analysis

For theory testing and public policy it is often important to relate the sensitive behavior to various characteristics of the participants, such as personality traits or sociodemographics. However, this is impossible if only a single question is asked (as in models M_1 to M_4), except in simple cases such as when comparing two different samples on the single item. In other words, inferences at the individual level are impossible when a single item is asked because, by definition, the individual's anonymity is guaranteed at that single item. Single items are also fallible and are only partial indicators of constructs, such as the tendency to engage in risky sexual behavior. This compromises the validity and reliability of inferences. To improve on this, multiple items are required to measure the underlying construct of risky sexual behavior. To more fully capture the underlying construct of risky sexual behavior tendencies, another item could be added, such as "During the last six months, how often did you have anal sex with a casual partner?" Still, having multiple items, by itself, does not suffice to enable individual-level analyses because the randomized response to each of the items protects the individual.

A recent innovation solved this shortcoming by combining RR procedures during data collection with a statistical IRT model during data analysis (Böckenholt & van der Heijden, 2007; de Jong et al., 2010; Fox, 2005; Fox & Meijer, 2008; Fox & Wyrick, 2008). This combination forms the core of IRRT. Because the responses to the individual items are randomized during data collection, the anonymity of the participants is protected at each specific item in the set. However, the application of an IRT model during the statistical analysis makes it possible to determine the scores of each participant on the underlying latent construct, and these scores can be related to other variables of interest.

For example, consider the situation where a number of binary items measure risky sexual behavior and a coin is used as a randomization device. Individuals indicate their true answer in case of heads and always answer *yes* in case of tails. Imagine that one observes a particular individual responding *yes* to all items. The likelihood is very low that all coin flips result in tails coming up. So, an individual with many *yes* answers is more likely to actually engage in some of the risky sexual behaviors than an individual with many *no* answers, even though it is not known which specific behaviors. In addition, some risky sexual behaviors are actually engaged in more often in the sample than are other risky sexual behaviors, and the questions about these former behaviors should have a higher reported proportion of *yes* answers across the sample of individuals. This information can be used to order the items on the underlying construct, thus making it possible to estimate the probability that an individual with a specific response pattern has actually engaged in a specific risky sexual behavior. In this way, IRRT can establish the true scores of individuals on a latent construct from randomized responses to multiple questions while still preserving anonymity about the response to each of the specific questions. The scores of individuals on the latent construct can then be related to other variables.

During the statistical analyses of the data, an IRT model is used to determine the latent constructs and the scores of the participants on them. Fraley et al. (2000) provide a nontechnical description of the basic IRT idea (see also Segura & González-Romá, 2003; L. L. Smith & Reise, 1998; Tuerlinckx, De Boeck, & Lens, 2002). A latent construct (or latent variable) is commonly designated by the letter theta (θ). In this case, it represents the latent tendency to engage in risky sexual behavior. Theta is often assumed to be normally distributed with a mean of 0 and variance of 1, that is, $\theta \sim N(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 1$. Together with item parameters (discrimination and difficulty parameters), the latent construct determines a so-called item characteristic curve. The latter describes the relationship between the latent construct and the probability that individuals give a particular response to a question. Figure 2 shows the typical situation for which IRT models are used. Note that there is no randomized response procedure in Figure 2; it is assumed that questions are administered directly. In the figure, the single latent variable θ influences each of K items. In addition, there is measurement error (indicated by the small arrows pointing to the observed scores) that causes the relationship between the latent variable and the item score to be probabilistic. Indexing individuals with i, the item characteristic curve for polytomous items if they are administered directly (i.e., without randomization) is (Samejima, 1969)

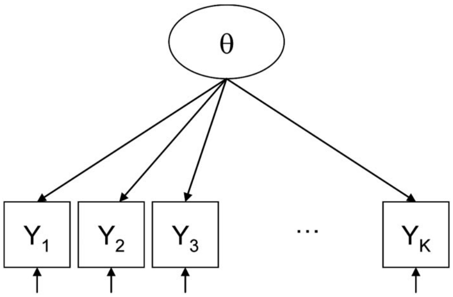

Figure 2. Item response theory (RT) measurement model.

$$P(Y_{ik} = c) = 1/[1 + \exp(-\alpha_k(\theta_i - \beta_{k,c-1})] - 1/[1 + \exp(-\alpha_k(\theta_i - \beta_{k,c})] \quad (1)$$

Y_{ik} denotes the observed response of individual i on item k (in this case, $c = 1$ would correspond to a *Never* answer, and $c = 5$ would correspond to an *Often* answer); θ_i is individual i's latent risky sexual behavior tendency score. The item parameter α_k indicates how well items discriminate among individuals ranking high and low on the latent variable, and it is usually above .50. It is conceptually similar to a factor loading. The item parameters $\beta_{k,c}$ are called difficulty parameters (for a 5-point response scale, there are four difficulty parameters). Figure 3 illustrates category probability functions for the parameter values (α_k, $\beta_{k,1}$, $\beta_{k,2}$, $\beta_{k,3}$, $\beta_{k,4}$) = (1.50, −1.50, −0.30, 0.90, 2.10).

It can be seen that the parameter $\beta_{k,c}$ represents the point where the probability is .50 that the item response is greater than option c. Furthermore, the probability of a response "1" (*Never*) becomes smaller and smaller as the latent variable increases, while the converse is true for a "5" (*Often*) response. The other response options rise and fall in the intermediate ranges of theta. For more details, see Fraley et al. (2000).

Model M_6 in Table 1 was obtained by combining this IRT model with polytomous randomized responses to multiple items (de Jong et al., 2010; Fox & Wyrick, 2008). Model M_5 was a special case. Thus, given the idiosynchrasies of single items, using multiple items serves better the twin goals of allowing individual-level inferences and estimating more general construct scores. The interest is in the parameter θ_i (risky sexual behavior tendency that can be related to other variables) and in the item parameters α_k and $\beta_{k,c}$. Moreover, if multiple items are used, nonadherence to the randomized response procedure can be dealt with in a natural way by specifying two distinct groups (or classes) of participants. The first group of participants will choose the socially safe response (*no* or *never* in the example) no matter what, whereas the second group adheres to the randomized response procedure (Böckenholt & van der Heijden, 2007; de Jong et al., 2010). Participants belong to the nonadherence class with a certain probability κ and to the adherence class with probability $1 - \kappa$, yielding Models M_7 and M_8.

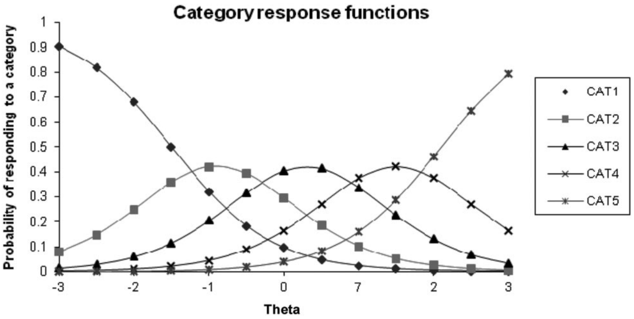

Figure 3. Item characteristic curves graded response model.

Multigroup Comparisons

For cross-cultural and other comparative research, it is important to be able to handle multigroup comparisons (Chen, 2008; Wicherts, Dolan, & Hessen, 2005). Model M_9 (MIRRT) is presented to deal with such situations. Because the samples in the empirical studies, which are described later, are cross-national, participant *i* in country *j* is indicated as *ij* (countries will be indexed by $j = 1, \ldots, J$). In the model, the latent variable θ_{ij} is assumed to have a population/country specific mean μ_θ^j and variance $\sigma_\theta^{2,j}$. For identification, the mean and variance for one of the countries have to be fixed (Reise, Widaman, & Pugh, 1993). The mean of the latent variable is set equal to 0 in this benchmark country, and the variance equal to 1. Finally, a latent class for nonadhering participants can be added, yielding Model M_{10}.

The item parameters are initially set as common across countries. This implies the assumption of measurement invariance. Measurement invariance refers to "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, p. 117). Measurement is fully invariant if the relations between all trait indicators and the trait of interest (i.e., the item parameters) are the same across countries. Before making comparisons across countries and interpreting item parameters, psychometricians often test the assumption of measurement invariance. If full measurement invariance does not hold, one can proceed to specify a model with partial measurement invariance. Partial measurement invariance implies that a subset of items operates similarly across countries but that another subset may not. Appendix B provides the details. It should be noted that tests for measurement invariance in IRT models with a large number of countries are difficult and still underdeveloped.

In summary, the MIRRT model applies both RR procedures during data collection on multiple items each with polytomous responses and IRT to infer the true scores of people on a latent construct of sensitive behavior, while accounting for people who do not adhere to the randomized response procedure. These true scores can be compared across countries and related to sociodemographic and other characteristics that are of theoretical and practical interest. Two cross-national studies of sexual attitudes and behaviors demonstrate the implications of these features.

Estimation

Bayesian inference is used to estimate the MIRRT model. A Bayesian model requires setting up a full probability model, that is, a joint probability distribution for all observable and unobservable quantities. Next, the posterior distribution of the model parameters given the data is computed. Often, a Markov chain algorithm is used to sample from the conditional posterior distributions (Gelman, Carlin, Stern, & Rubin, 2004). For those using a statistical software program such as WinBugs (Lunn et al., 2000), Bayesian analysis is relatively easy and requires only a few lines of code. It is not necessary for psychologists to derive the full conditional distributions, because WinBugs does this. The more technically inclined researchers can find the full conditional distributions of the IRRT model in other articles (e.g., Fox & Wyrick, 2008).

In contrast, the traditional maximum likelihood estimation of MIRRT and similar methods requires a complex EM algorithm with optimization routines and derivatives. For estimation of the parameters of interest, this study used 30,000 burn-in iterations, and the last 10,000 iterations were used for calibration, that is, to compute posterior quantities of interest (such as posterior means and standard deviations). Henceforth, significance is indicated using terminology derived from frequentist statistics. If 0 is not included in the 95% posterior credible interval of the parameter, a parameter is considered "significant at 5%."

To test for mean differences between countries, it is not possible to use the standard *F* test derived from analysis of variance. The reason is that the model is cast in a Bayesian framework that necessitates a Bayesian approach to hypothesis testing. A Bayesian *F* test can be computed with formulas presented in Box and Tiao (1973). De Jong, Steenkamp, and Fox (2007) presented an application in cross-cultural research. Appendix C provides the derivation of the Bayesian *F* test.

Model fit is not yet well developed for IRRT models. The confirmatory factor analysis framework employs popular statistics such as root-mean-square error of approximation (RMSEA) or comparative fit index (CFI) to evaluate goodness of fit, but such universal statistics do not exist for IRT models. Despite the absence of universal fit statistics, it is possible to use posterior predictive checks in the Bayesian framework (Sinharay, Johnson, & Stern, 2006). This approach relies on a specific discrepancy measure. De Jong et al. (2010) used a Bayesian residual analysis to check model violations and defined some intuitive discrepancy measures. Nonetheless, more research is necessary to evaluate which discrepancy measures are most suited for checking model fit. In the current research, which relied on WinBugs for model estimation, we considered the magnitude of the discrimination parameters (they should be significant and larger than .5), as well as the reliability of the construct.

Study 1

Theoretical Background

Study 1 aimed to demonstrate how to apply MIRRT and how to interpret the output from an MIRRT analysis. Many personality and social psychologists are interested in understanding the relationship between personality and a variety of attitudes and behaviors. In Study 1, the substantive focus was on personality and sociodemographic correlates of permissive sexual attitudes and risky sexual behavior. Questions about sex are typically seen as sensitive, and scholars have raised concerns about using direct questions to collect information about sex (Catania et al., 1990). If direct questions are used, privacy concerns and shame would be two potent factors that may cause differences between true and reported attitudes or behavior. Study 1 applied MIRRT in a two-country Web survey and addressed the following research question:

> RQ1: What are the relationships between, respectively, permissive sexual attitudes and risky sexual behavior, and personality traits and sociodemographic characteristics of individuals?

It is reasonable to predict that permissive sexual attitudes are positively related to extraversion and openness to experience—two of the of the Big Five personality dimensions—and negatively related to conscientiousness.

Extraverts seek out sexual stimulation, which may lead to more interest in sexuality. Openness has been related to liberal value systems, need for variety in actions, and richness of fantasy life (McCrae & Costa, 1997). Individuals who score high on openness may thus develop a more permissive sexual outlook. Conscientiousness correlates with a disposition toward cautiousness and criticality, orderliness, tidiness, and rule compliance (Hogan & Ones, 1997). Conscientious individuals may therefore hold less permissive attitudes. In support of this, Heaven, Fitzpatrick, Craig, Kelly, and Sebar (2000) found a positive relationship between extraversion and sexual curiosity and excitement using direct questioning in an Australian sample. Lameiras Fernández and Rodríguez Castro (2003) found in a sample of Spanish college students a positive relationship between permissive sexual attitudes and openness to experience and a negative relationship between conscientiousness and permissive sexual attitudes.

There is reason to believe that performing risky sexual behaviors is positively related to extraversion and negatively related to conscientiousness and agreeableness (Eysenck & Eysenck, 1975; Hoyle, Fejfar, & Miller, 2000; Schenk & Pfrang, 1986; Trobst et al., 2000). It is thought that a higher libido or a higher desired level of arousal prompts extraverts to seek out risky sexual encounters. In a large-scale study, Schmitt (2004) indeed observed a positive relationship between extraversion and sexual promiscuity, especially in western Europe. Given an inclination toward cautiousness, highly conscientious individuals are likely to reduce their exposure to sexual risk taking. Furthermore, among the Big Five, the agreeableness dimension is most concerned with interpersonal relationships. Frequent risky sexual behaviors with a casual partner do not reflect intimacy and may lead to mistrust and unstable relationships. This suggests a negative association between agreeableness and risky sexual behavior.

Prior research (e.g., Anderson & Dahlberg 1992; de Jong et al., 2010; Schmitt, 2003) reported permissive sexual attitudes and risky sexual behaviors to be more positive and higher among men and younger people than among women and older people. To the extent that higher education levels promote more openness, it is reasonable to expect that higher education is associated with more positive permissive sexual attitudes. Study 1 tested these predictions.

Method

Participants. SSI (Survey Sampling International), a professional survey organization that maintains cross-national online (web) panels, collected the data in 2011. Online surveys were purposely used because this impersonal data collection method (compared to telephone and face-to-face interviewing) reduced the likelihood of socially desirable responding and, combined with IRRT, increased validity (Tourangeau & Yan, 2007). SSI recruited 927 participants, 465 in Spain and 462 in the Netherlands. Age of the participants varied from 17 to 67 years, with an average of 33 in both countries. A unique feature was that both samples reflected the entire population, in contrast to samples that exclusively consist of students. In each country, approximately 50% of the participants were female.

Measures. To measure permissive sexual attitudes, participants completed the “permissiveness” component of the Brief Sexual Attitudes Scale (Hendrick, Hendrick, & Reich, 2006). This component consisted of 10 items. The scale has been shown to have good psychometric properties and to relate to other theoretically relevant constructs, such as relationship satisfaction, commitment, and respect for the partner. Risky sexual behaviors were measured using the corresponding component of Turchik and Garske’s (2009) sexual risk taking measure. This validated component consisted of five items and has been applied to reliably measure college students’ risky sexual behaviors.

Several person-level characteristics were used to examine the nomological network—the body of interlocking evidence that supports the validity of a construct—of permissive sexual attitudes and risky sexual behavior and to illustrate the strengths of MIRRT. First, sociodemographic information was collected. Participants listed their age (in years), gender (1 = female, 0 = male), and educational level (response options: “No formal education,” “education up to age 12,” “education up to age 14,” “education up to age 16,” “education up to age 18,” “higher education,” “university”). The questionnaire also asked participants about their sexual behavior orientation using the following item: “If you consider your sexual history, would you classify your sexual *behavior* as primarily homosexual, heterosexual, or bisexual? Note: sexual behavior includes passionate kissing, fondling, petting, oral-to-anal stimulation, hand-to-genital stimulation, vaginal, oral, and anal sex.” This item had three response options: “Homosexual,” “Heterosexual,” or “Bisexual.”

The Big Five personality dimensions were measured with Rammstedt and John’s (2007) short-form Big Five Inventory (BFI-10), which has been used before in cross-national research. In addition, self-ascribed conservatism was measured with an item with five response options from “extremely liberal” to “extremely conservative” (Eagly, Diekman, Johannesen-Schmidt, & Koenig, 2004). Conservatism, like the personality dimension openness to experience, relates to people’s value system that may express itself in sexual attitudes and behaviors.

All items tapping into sexual attitudes and behavior were administered via the polytomous RR procedure and operationalized using an electronic die, as illustrated in Figure 1. The instructions for permissive sexual attitude items were as follows:

> The next questions may be sensitive to you. Therefore, we want to protect your privacy by using a technique called “randomized response.” Your answers to the following questions will depend on the outcomes of several throws of an electronic die. We now explain how this works.
>
> The answer you give to a question depends on the outcome of the roll of the die, as follows:

> Procedure:
>
> Step 1: In order to answer a particular question you roll the electronic die for that question.
>
> Step 2: If the outcome of the first die roll is 1, 2, 3, or 4: Please give your honest and true answer to the question.
>
> Step 3: If the outcome of the first die roll is 5 or 6: Please roll the die again.
>
> Step 4: If the outcome of the second die roll is 1: Give answer 1 (Strongly disagree)
>
> Step 5: If the outcome of the second die roll is 2: Give answer 2 (Disagree)
>
> Step 6: If the outcome of the second die roll is 3: Give answer 3 (Neither agree nor disagree)
>
> Step 7: If the outcome of the second die roll is 4: Give answer 4 (Agree)
>
> Step 8: If the outcome of the second die roll is 5 or 6: Give answer 5 (Strongly agree)
>
> The procedure is repeated for each question. The flowchart in Figure 1 summarizes this procedure.
>
> Please take some time to study the procedure so that you understand it. It is important that for each question you follow the procedure outlined in the following illustration exactly (note: you have to throw the die at least once for each question), even if you don't find it difficult to give an honest answer to a question, or if the roll of the die produces a number that you do not like. The illustration indicates which answer to give. That is, for each question you either give your true and honest answer, or you give the forced answer that the die tells you.
>
> The idea behind this procedure is that only you know the outcomes of throwing the die. The outcomes of the throws of the die are not stored on the computer. Thus, it is impossible for us to retrieve your true answer to each question. However, because the frequencies of the outcomes of throwing a die are known (each outcome has the same probability, namely 1 out of 6), we can still determine what the true answer to a particular question is *across all participants.* Thus, we can make inferences at the sample level, but not at the individual level. Because this procedure fully protects your privacy you can, with all your heart, provide your true answer to a question if the die tells you to do so.
>
> In the questions below, by the term "sex" we mean all forms of vaginal, oral or anal intercourse. Vaginal sex means insertion of a penis in the vagina. Oral sex means having sex by using the mouth and tongue to stimulate the genital area. Anal sex means inserting of a penis in the anus. When we refer to a "casual partner" we mean people other than a formal boyfriend, girlfriend, or spouse. Casual sex is sex with a person that one has no longer term relationship with.

The instructions for risky sexual behavior were identical except for the response labels. The scale labels for risky sexual behavior ranged from *Never* to *Often* (see Figure 1). When the electronic die indicated a forced choice (i.e., the first roll produced a "5" or a "6"), the computer automatically filled in the answer after the second throw, and participants could not change that answer anymore. Hence, there was no opportunity for participants to alter their response if they did not like the outcome of the die or found it difficult to select the answer indicated by the die. Under this setup it is not possible to have a pattern of fully self-protective answers (i.e., a pattern with only "1" answers). Model M_9 in Table 1 was used to analyze the data of this study.

Results

The Netherlands was set as the benchmark country. Thus, the mean of theta was equal to 0 and its variance was equal to 1. In Spain, the mean and variance of theta were estimated freely. Separate models were estimated for permissive sexual attitudes and risky sexual behavior. The tests for measurement invariance indicated that a model with fully invariant item parameters had to be rejected for permissive sexual attitudes. A model with freely varying item parameters had a log-likelihood of −15,484, and a model with fully invariant item parameters had a log-likelihood of −15,580. The difference $-2 * \Delta LL = 191$ was significant given a difference in the degrees of freedom of 50 (5 * 10). Therefore, a model was specified with partial measurement invariance. Further analyses indicated that Item 5 was invariant, but the other parameters were better left unconstrained. An interesting finding was that the mean difference between the Netherlands and Spain in permissive sexual attitudes was larger if full invariance was imposed than if only partial measurement invariance was imposed. The reduction in the mean difference parameter by specifying a model with only partial measurement invariance was almost 50%. The item parameters thus absorbed some of the differences between countries that would have otherwise been (unjustly) attributed to country differences in the latent construct.

Invariant item parameters were supported for risky sexual behavior. The log-likelihood for a benchmark model with noninvariant item parameters was −5,966, compared to a log-likelihood of −5,985 for a model with invariant item parameters. The LR-statistic $-2 * \Delta LL = 39$ was not significant, given a difference in the degrees of freedom of 25 (5 * 5). One reason, albeit speculative, for the contrast with the attitudinal data was that the risky sexual behavior items were very specific and detailed, which may have led to a higher degree of equivalence across countries. The analysis for risky sexual behavior, therefore, proceeded with invariant item parameters.

Item parameters. The item parameters for permissive sexual attitudes are in Table 2, and the item parameters for risky sexual behavior are in Table 3. The discrimination parameters for the permissive sexual attitudes items in the Netherlands ranged from 1.44 to 4.32 and suggested that all items provided adequate discrimination (i.e., they were significant and larger than .5). The discrimination parameters of the noninvariant items were generally a bit lower in Spain, even though they were still significant.

The threshold parameters are sometimes called *difficulty* parameters and are on the same scale as the latent trait. The larger the first threshold, the more difficult it was for a participant to have a higher score on the item for a given theta. Thus, in order for a person to "have sex with someone that I do not like very much" ($\beta_{6,1} = 0.38$ in the Netherlands), the person needed a higher theta score than to "have sex with someone without having a long-term relationship with that person" ($\beta_{1,1} = -0.82$). The most difficult items in the Netherlands were Items 3 ("would like to have sex with many partners") and 9 ("sex with a person that I do not like very much"). In Spain, the most difficult item to endorse was Item 7 ("the best sex is when people have no attachments").

The discrimination parameters for risky sexual behavior items were significant. Especially Item 1 ("vaginal intercourse with casual partner without a latex condom") and Item 2 ("vaginal intercourse with a casual partner without protection against pregnancy") had larger discrimination parameters, which indicated that

Table 2
Operating Characteristics of Permissive Sexual Attitudes Items: Study 1

		Item characteristics									
		Discrimination		Threshold 1		Threshold 2		Threshold 3		Threshold 4	
No.	Item	NL	Spain	NL	Spain	NL	Spain	NL	Spain	NL	Spain
1	I can have sex with a person without having a long-term relationship with the person.	3.73	2.56	−0.82	−0.82	−0.35	−0.04	0.07	0.65	0.98	1.55
2	It is acceptable for me to have casual sex with someone.	3.16	3.16	−0.80	−0.80	−0.20	−0.20	0.34	0.34	1.36	1.36
3	I would like to have sex with many partners.	2.11	2.11	−0.08	−0.08	0.77	0.77	1.381	1.38	2.24	2.24
4	Casual sex for a single night is attractive to me.	4.32	1.92	−0.34	−0.94	0.22	−0.02	0.82	0.81	1.72	2.13
5	It is okay for me to have ongoing sexual relationships with more than one person during a certain period.	2.71	1.77	−0.29	−0.86	0.34	0.10	0.94	1.02	1.64	2.36
6	Sex as a simple exchange of favors is okay to me if both people agree to it.	1.80	1.65	−1.01	−0.69	−0.40	0.07	0.35	0.83	1.46	2.21
7	To me, the best sex is when people have no attachments.	2.03	1.01	−0.35	−0.05	0.22	1.39	1.23	3.23	2.28	3.68
8	My life would have fewer problems if I could have sex more freely.	2.19	1.29	−0.23	−0.69	0.60	0.30	1.41	1.59	2.07	2.90
9	I can enjoy sex with a person that I do not like very much.	1.44	0.93	0.38	−0.32	1.14	1.30	1.92	2.66	3.71	5.16
10	It is okay for me to have sex as form of physical release.	2.50	1.30	−0.62	−1.90	−0.09	−0.97	0.50	0.30	1.71	1.75

Note. NL = the Netherlands.

these behaviors were most strongly related to the latent risky sexual behavior construct. The threshold parameters were also quite similar, which suggested that most of the risky sexual behavior items were of similar difficulty.

Even though the measurement scales used have been shown to display adequate reliability in prior research, the item parameters can be used to compute the construct reliability for permissive sexual attitudes and risky sexual behavior. We calculated reliability through the item information functions (Fraley et al., 2000). This is plotted in Figure 4. Reliability for IRT models was not constant over the range of the construct, as seen in Figure 4. Along most of the trait range, reliability was slightly higher for permissive sexual attitudes than for risky sexual behavior. One reason for this was that there were more items to measure permissive sexual attitudes. It can also be seen that reliability was less high for respondents low on risky sexual behavior.

Associations with person-level characteristics. A strength of MIRRT is that latent construct scores can be related to individual-level covariates. Columns 2 and 4 in Table 4 relate both permissive sexual attitudes and risky sexual behavior scores to

Table 3
Operating Characteristics of Risky Sexual Behaviors Items: Study 1

		Item characteristics				
No.	During the last 12 months (counting from today), how often did you . . .	Discrimination	Threshold 1	Threshold 2	Threshold 3	Threshold 4
1	Have vaginal intercourse with a casual partner without using a latex or polyurethane condom? (Note: include times when you have used a lambskin or membrane condom)	4.28	0.62	1.09	1.72	2.27
2	Have vaginal intercourse with a casual partner without protection against pregnancy?	3.04	0.67	1.20	1.99	2.39
3	Given to or received fellatio (oral sex on a man) from a casual partner without a condom?	1.71	0.49	0.87	1.67	2.42
4	Given to or received cunnilingus (oral sex on a woman) from a casual partner without a “dental dam” (or other adequate protection)?	1.63	0.49	0.86	1.63	2.56
5	Have sex with a casual partner while you or your casual partner used alcohol or drugs before or during sex?	1.72	0.59	1.03	2.00	2.89

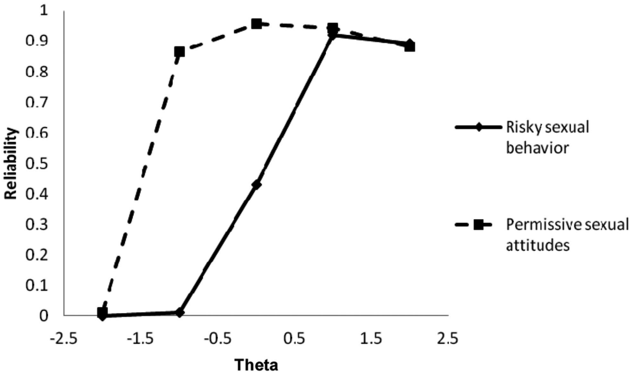

Figure 4. Reliability for permissive sexual attitudes and risky sexual behavior.

sociodemographic variables. A country dummy variable (1 = Spain) was included in the regression analyses to control for intercept differences. Spanish participants scored higher on permissive sexual attitudes and risky sexual behaviors.

Women generally scored lower on permissive sexual attitudes and risky sexual behavior. The gender effect for permissive sexual attitudes was consistent with findings of gender differences in sociosexuality by Schmitt (2003, 2005). The effect of education was positive and significant for permissive sexual attitudes but insignificant for risky sexual behavior.

Columns 3 and 5 present the estimated parameters for models that contained sociodemographic variables, Big Five personality dimensions, and the conservatism measure. Permissive sexual attitudes related primarily to openness to experience and conscientiousness.

The results for risky sexual behavior did not mirror those for permissive sexual attitudes. That is, although openness was related to permissive sexual attitudes, it was not significantly associated with risky sexual behavior. Consistent with prior literature (Hoyle et al., 2000; Trobst et al., 2000), conscientiousness and agreeableness were negatively related to risky sexual behavior.

Conservatism, like openness to experience, related to the individual's value system. Permissive individuals were less likely to accept traditional moral views and were more comfortable with the ideas of sexual gratification and same-gender sex. This explained the negative relationship between conservatism and permissive sexual attitudes. However, having a more permissive value system did not necessarily relate to more risky sexual behavior, as can be seen from the nonsignificant association of conservatism with risky sexual behavior.

The risky sexual behavior measure contained two items that specifically described vaginal intercourse, which then raised the question whether an analysis across all participants, including those engaging in same-gender sexual behavior, was appropriate for that measure. To examine this issue, we conducted a follow-up analysis on participants who reported that their past sexual behavior was exclusively heterosexual. The patterns of results (significance, valence) for the initial and follow-up analyses were identical, which indicated robustness of the results with respect to sexual behavior orientation.

Discussion of Results

The findings for permissive sexual attitudes were in line with Lameiras Fernández & Rodríguez Castro (2003). Openness to experience and conscientiousness were most strongly related to permissive sexual attitudes. Furthermore, men, more educated individuals, and individuals low on conservatism reported more permissive sexual attitudes.

The results for risky sexual behavior were more interesting. Although the patterns for conscientiousness and agreeableness were in line with prior research, a somewhat surprising finding of this study was inconsistent with previous research: extraversion was unrelated to risky sexual behavior. Several studies found that extraverts engage in more risky sex than do introverts (Eysenck & Eysenck, 1975; Schenk & Pfrang, 1986; Schmitt, 2004). In a large-scale study, Schmitt (2004) found a positive relationship between extraversion and sexual promiscuity, especially in western Europe. A possible speculative explanation was that extraverts admitted such behavior more easily than introverts, creating a significant effect of extraversion in research using direct questions. Under conditions of anonymity, as in the

Table 4
Regression Analyses for Permissive Sexual Attitudes and Risky Sexual Behavior: Study 1

	Attitudes		Behaviors	
Variable	Permissive sexual attitudes	Permissive sexual attitudes	Risky sexual behaviors	Risky sexual behaviors
Constant	−0.04 (.20)	−0.02 (.39)	0.39 (.18)	0.81 (.30)***
Country (1 = Spain)	0.37 (.06)***	0.43 (.07)***	0.35 (.05)***	0.37 (.05)***
Age	0.00 (.01)	0.01 (.01)	−0.01 (.01)	−0.01 (.01)
Gender (1 = female)	−0.68 (.06)***	−0.75 (.07)***	−0.37 (.05)***	−0.33 (.05)***
Education	0.06 (.03)***	0.07 (.03)***	−0.01 (.02)	0.01 (.02)
Big Five				
Openness to Experience		0.21 (.04)***		0.03 (.04)
Conscientiousness		−0.14 (.04)***		−0.13 (.03)***
Extraversion		0.03 (.04)		0.02 (.03)
Agreeableness		−0.03 (.04)		−0.09 (.04)***
Neuroticism		−0.04 (.04)		0.03 (.03)
Conservatism		−0.39 (.03)***		0.02 (.02)
R^2	17.7%	24.2%	10.1%	13.1%

Note. Unstandardized regression coefficients shown.
*** $p < .01$.

present study, such reporting differences may have vanished, nullifying the effect of extraversion on risky sexual behavior.

The significant gender effect indicated that men displayed more risky sexual behavior than women. Education and conservatism, however, did not have an effect. Thus, less conservative individuals and those with higher education reported more permissive attitudes but did not engage more in risky sexual behaviors.

Study 2

Theoretical Background

Study 1 presented a two-country MIRRT analysis of permissive sexual attitudes and risky sexual behavior. It was conducted with an electronic die that prevented survey participants from not adhering to the randomization device and focused on person-level characteristics such as personality dimensions. Study 2 generalized and extended Study 1. Study 2 focused on permissive sexual behavior and applied MIRRT to a much larger set of countries, which allowed relating the behavior to person-level as well as country-level characteristics. In addition, Study 2 used a real rather than an electronic die to allow for nonadherence to the randomized response procedure. Model M_{10} was estimated to accommodate this data collection feature (see Table 1). More specifically, Study 2 addressed the following research questions:

> RQ2: What is the relationship between permissive sexual behavior and person-level characteristics, such as gender and age?
>
> RQ3: What is the relationship between permissive sexual behavior and country-level characteristics (e.g., sociocultural variables)?
>
> RQ4: How do the relationships between permissive sexual behavior and person-level characteristics vary across countries?

In order to examine the research questions, Study 2 compared 17 countries on permissive sexual behavior. The countries represented a diverse array of geographic, cultural, and linguistic categories. The study followed up on recent programs of research that collected multicountry data to document cross-cultural differences in personality and behavior (Costa, Terracciano, & McCrae, 2001; Diener & Diener, 1995; McCrae & Terracciano, 2005; Schmitt, 2003, 2005; Schwartz & Rubel-Lifschitz, 2009).

Which predictions can be made about gender effects, age effects, age-by-gender effects, and cultural effects? The first prediction is that men will display higher levels of permissive sexual behavior than women. Indeed, many studies found such gender differences in promiscuous sexual behavior, both within and across countries (Anderson & Dahlberg, 1992; Buss & Schmitt, 1993; Li & Kenrick, 2006; Schmitt, 2003, 2005). Scholars have proposed both evolutionary (e.g., parental investment theory or sexual strategies theory (Buss & Schmitt, 1993) and social structural explanations for these differences (Eagly & Wood, 1999).

The second prediction is that younger age groups will display more permissive sexual behavior than older age groups. This is to be expected in view of general value changes across age cohorts and the human sexual cycle. In younger age groups, sexual urges are still strong, and stable marital and other intimate relationships may not have fully formed yet. Younger people are also more likely to engage in short-term sexual activity because it provides information about one's own mate value and about the quality of potential mates. The older age groups will generally display the lowest incidence due to increasingly stable relationships, decreasing sexual desires, and changing sexual anatomy and physiology (Kennedy, Martinez, & Garo, 2010; Kenrick et al., 1995).

The third prediction is that permissive sexual behavior will vary across cultures as a function of sociocultural variables. In his landmark study, Schmitt (2005) found that national differences in sociosexuality could be predicted from the sex ratio and the level of environmental demand in a country. The sex ratio in a nation is defined as the balance of marriage-aged men versus marriage-aged women (Pedersen, 1991). Countries with lower sex ratio (more women than men) are expected to score higher on permissive sexual behavior. There are various possible theoretical mechanisms that lead to this expectation. A reason proposed by evolutionary psychologists is that competition between women for men increases under such conditions. Conversely, Guttentag and Secord (1983) proposed, in line with social structural theory, that sex ratios affect the values of the social exchanges between men and women in relationships (Eagly & Wood, 2005). Nonetheless, both mechanisms would lead to the same prediction.

Demanding environments are those in which there are fewer resources and a higher level of stress. It is more difficult to rear offspring in such environments. The sociocultural variable *environmental demand* is a building block for strategic pluralism theory (Gangestad & Simpson, 2000). The theory predicts that in demanding environments, permissive sexual behavior will be lower because the adaptive need for biparental care and heavy family investment increases.

The fourth prediction is about the size of gender effects across various age groups. The unique sampling frame of Study 2 permitted an analysis of gender differences in permissive sexual behavior across age groups. Prior studies on gender effects have often used student participants, and it is relevant to know whether gender effects on permissive sexual behavior found among students hold up in society at large (Henrich, Heine, & Norenzayan, 2010). Nearly all the samples in the International Sexuality Description Project (ISDP; Schmitt, 2005) are based on college students. The actual incidence of permissive sexual behavior across age groups is an empirical question that, to the best of our knowledge, has not been systematically examined in prior research.

A final yet pivotal issue addressed by research question 4 is whether observed gender differences have an evolutionary basis (Schmitt, 2005), or whether they derive from different socialization and social structure (Eagly & Wood, 1999). The variation of gender effects across age groups and countries may inform the debate between evolutionary and social structural source theories of gender differences, although definite answers will require much more work.

Social structural theory predicts that changes in division of labor, values and roles of men and women, and improved contraceptive and medical technology will cause a smaller gender difference among younger participants than among older participants. It also predicts that gender differences will be larger in traditional cultures where women are more constrained (measured by the level of women's development; see the Method section). In contrast, evolutionary theories, such as strategic pluralism theory, predict that the size of gender differences should decrease as environments become less demanding. The results in Schmitt (2005) only marginally supported evolutionary theory, as none of the variables that indicated high environmental demand was directly correlated with the magnitude of the gender difference. The results seemed more in line with social structural theory.

Method

Participants. As in Study 1, SSI collected the data by means of a Web survey. Data collection took place in 2008 across 17 nations and gender (male–female) and age groups (three almost equally large groups: < 41 years, 41–55 years, and > 55 years). Data collection was the same as in Study 1, except that participants used a regular die to respond to the questions.

Sample sizes were as follows: Belgium (n = 437), Brazil (n = 400), Canada (n = 334), Denmark (n = 394), Estonia (n = 255), France (n = 408), Germany (n = 365), India (n = 316), Italy (n = 376), Japan (n = 347), Netherlands (n = 404), Poland (n = 393), Portugal (n = 319), Singapore (n = 320), Switzerland (n = 346), UK (n = 348), and US (n = 433). The total sample size was 6,195, with 50.3% of the sample being female. Of total participants, 33.2% were under the age of 41, 35.8% were between the ages of 41 and 65 years, and 31.0% were over the age of 65 years. Table 5 provides a three-way breakdown of the sample in age-by-gender-by-country. The US was set as the baseline country (and hence, the mean and variance in the US are fixed to 0 and 1 respectively in the analyses).

Measure. The self-report measure of permissive sexual behavior had four items. The measure used in this study was part of a large-scale project designed by a research center of a large western European university. Items are listed in Table 6. The original items were formulated in English. The items were slightly adapted after pretesting in the US (n = 528) so that they displayed good reliability. Next, they were translated by bilinguals into 10 languages, namely, Danish, Dutch, English, Estonian, French, German, Italian, Japanese, Polish, and Portuguese. These were the languages commonly spoken by native speakers in the 17 sampled countries. For countries or regions that officially shared the same native language but with dialect variations (e.g., French in Belgium or Quebec compared to French in France), the English version of the questionnaire was also translated by a native speaker from that particular country or region to ensure the terms retained the same meaning.[1]

Participants first read: "In the questions below, except if explicitly mentioned otherwise, by sex we mean vaginal, oral or anal intercourse. When we refer to your steady partner we mean your boyfriend, girlfriend, or spouse only. When we refer to a nonsteady partner we mean people other than your boyfriend, girlfriend, or spouse." This definition of sex was consistent with other research (e.g., Weinhardt et al., 1998) and ensured that participants understood the terminology in the items.[2]

Each item was introduced as "During the last six months how often did you have . . . ," followed by "sex with a nonsteady partner" (Item 1), "unplanned sex with a nonsteady partner" (Item 2), or "sex under the influence of alcohol" (Item 3). The fourth item had an additional instruction: "Some people have, at some point in their life, two or more sexual partners (this does not necessarily mean a threesome or having sex with multiple people at the same time)." After this instruction, the item was "During the last six months how often did you maintain sexual relations with two or more partners at the same time?" The 5-point response scale for the four items was *Never*, *Rarely*, *Sometimes*, *Often*, and *Very often*.

Country-level characteristics. Table 7 lists the sociocultural characteristics used at the country level. National *sex ratios* were obtained from the United Nations Statistics Division (2001), as in Schmitt (2005). The level of *environmental demand* was measured by several variables. Fertility rates, GDP/capita, and the Human Development Index (the achievement of a nation in basic human capabilities, including health, longevity, education, and a decent standard of living) were extracted from the United Nations Development Programme (2001). The UNICEF Global Database was used to obtain the percentage of low birth weight infants. The mean age of women at marriage was obtained from the World's Women 2005 Report (United Nations Statistics Division, 2001). Finally, the United Nations Statistics Division provided indices for infant mortality rates.

Testing social structural accounts of gender differences required variables that measure the level of *women's development*. Several women's development indexes were considered, such as the number of women in parliament (United Nations Statistics Division, 2001), gender empowerment (United Nations Development Programme, 2001), and the percentage of female-headed households (United Nations Statistics Division, 2001). Finally, the level of contraceptive use, which may shed more light on the validity of social structural theory, was also extracted from the United Nations Statistics Division.

Results

Item parameters. The posterior means of the item parameters are listed in Table 5.[3] These estimates provided information about the discriminating power of the specific items, as well as their difficulty values. All items were discriminating (i.e., they were larger than .5). Item 2 on unplanned sex with a nonsteady partner was most discriminating. Turning to the threshold parameters, it is clear that sex under the influence of alcohol (Item 3) was relatively common. That is, participants did not need a high score on the latent construct in order to pass the first response option for this item. However, the alcohol item had high threshold values for response options 3 and 4. In effect, these data illustrate Lord's paradox: People admit more commonly to have sex under the influence of alcohol than with an nonsteady or new partner (it is easier to get alcohol than a new or nonsteady partner), but they typically do not do it very often, which accounts for the crossover.[4]

The category response functions provided the "linking mechanism" between item parameters and observed scores. Figure 5

[1] The measure of sexual behavior originally contained an item intended to measure condom use during intercourse. We decided to drop this item because it reduced reliability of the scale in a large pilot test (n = 2,500) across the 17 countries, and it was inconsistently completed across countries, perhaps due to the different implications for men and women and age groups.

[2] In the US and the UK, two of the items had a suboptimal translation. That is, the term "unsteady partner" was used instead of "non-steady partner," which could be interpreted as someone who is "emotionally unbalanced" or who is "physically shaky." Because the instructions in these two countries did define the meaning of the term, the threat of an ambiguous interpretation was somewhat mitigated.

[3] Testing for measurement invariance in IRT models with a large number of countries is difficult. Hence, we did not do this. One approach in psychometrics (de Jong et al., 2007) is to impose a random-effects prior for discrimination and threshold parameters. However, this is impossible in WinBugs; moreover, the random-effects prior has so far not been integrated with randomized response models in the literature.

[4] We thank an anonymous reviewer for this suggestion.

Table 5
Three-Way Breakdown of the Number of Participants by Age, Gender, and Country: Study 2

	Age < 41		Age 41–55		Age > 55	
Country	Male	Female	Male	Female	Male	Female
Belgium	49	71	83	67	87	80
Brazil	76	94	99	87	21	23
Canada	66	80	64	38	40	46
Denmark	39	65	68	82	67	73
Estonia	37	84	27	52	16	39
France	55	83	62	78	61	69
Germany	63	70	97	39	43	53
India	153	116	28	14	4	1
Italy	88	76	48	52	57	55
Japan	44	49	60	64	71	59
Netherlands	53	71	72	77	66	65
Poland	71	156	82	63	15	6
Portugal	89	92	48	51	15	24
Singapore	102	106	45	50	8	9
Switzerland	35	76	81	76	43	35
United Kingdom	32	54	80	53	73	56
United States	121	79	102	48	60	23

shows the category response functions for Item 3 (sex under the influence of alcohol), derived from the item parameters in Table 5. For each value of theta, the corresponding probabilities of a particular response could be derived from the graph. For instance, for $\theta = 0$ (the mean for permissive sex in the US), the probability of a *Never* response was 0.31 (thus, given the model, 31% of U.S. participants had actually never had sex under the influence of alcohol), the probability of a *Rarely* response was 0.28, the probability of a *Sometimes* response was 0.30, the probability of a *Regularly* response was 0.08, and the probability of an *Often* response was 0.03. In India, the mean for permissive sex was $\theta = -1.23$ (see Table 4), which resulted in the five response option probabilities [Pr(1), Pr(2), Pr(3), Pr(4), Pr(5)] = [0.69, 0.19, 0.10, 0.02, 0.00], where Pr(c) indicated the probability that $Y = c$. The probability of a *Never* response was clearly much higher for an "average participant" in India than in the US. Reliability curves were not calculated in Study 2 because the large number of countries made pooling of the information curves statistically suspect and 17 plots (one for each country) cumbersome. However, reliability of the measure was good in a pretest that was conducted in several pilot countries.

Procedural adherence. Table 8 lists the posterior mean percentage of participants who did not adhere to the procedure. There were important cross-national differences in nonadherence, $\chi^2(16) = 255.8$, $p < .01$. For instance, Japanese and Brazilian participants adhered to the procedure very well (only 5% nonadherence), whereas Indian participants were more likely to not follow the procedure (29% nonadherence). It emphasized the importance of accounting for nonadherence, in particular in cross-cultural research.

Gender differences. Figure 6 displays gender means in the countries for permissive sex. The graph is sorted by male mean, so that the country with the smallest mean for men is displayed left on the *x*-axis, and the country with the highest mean for men is displayed right on the *x*-axis. Higher scores on the *y*-axis denoted more frequent permissive sexual behavior. It should be noted that a score of 0 on the *y*-axis corresponded to the mean of permissive sexual behavior in the US. Thus, nearly all countries had a mean below 0, which implied a lower mean than the US. Men generally reported more permissive sexual behavior than women did. A series of *F* tests with gender as factor and permissive sexual behavior as dependent variable yielded significant differences at the 5% significance level for each country, except for Germany and Switzerland (details of the *F* tests for each of the countries are available from the authors). Thus, we found in nearly all countries a significantly higher mean for men than for women. Even in the two countries where insignificant differences were obtained, the male mean was always higher.

Table 6
Operating Characteristics of Permissive Sexual Behavior Items: Study 2

		Item characteristics				
No.	Item	Discrimination	Threshold 1	Threshold 2	Threshold 3	Threshold 4
1	Sex with nonsteady partner	5.42	0.11	0.78	1.50	2.28
2	Unplanned sex with nonsteady partner	6.11	0.03	0.81	1.65	2.37
3	Sex under influence of alcohol	1.29	−0.61	0.30	1.71	2.82
4	Sexual relations with two or more partners at the same time	1.37	−0.38	0.82	2.08	3.10

Table 7
Relationship of Sociocultural Variables With Permissive Sexual Behavior: Study 2

Variable	Permissive sexual behavior	Permissive sexual behavior gender gap
National sex ratio (women/men)	0.54**	−0.42**
Level of contraception	0.67***	−0.67***
Environmental demand		
Familial stress		
% low birth weight (n = 16)	−0.62***	0.45**
Infant mortality (n = 17)	−0.46**	0.37*
Economic resources		
GDP/capita (n = 17)	0.29	−0.44**
Human Development Index (n = 17)	0.41**	−0.39*
Mortality		
Life expectancy (n = 17)	0.16	−0.22
Prolific reproduction		
Fertility rates (n = 17)	−0.25	0.27
Mean age at marriage (n = 16)	0.16	−0.24
Women's development		
% of women in parliament (n = 17)	0.25	−0.35*
Gender empowerment (n = 16)	0.09	−0.43**
% of women-headed households (n = 13)	0.77***	−0.58**

Note. GDP = gross domestic product.
* $p < .10$. ** $p < .05$. *** $p < .01$.

Table 8
Summary of Results for Permissive Sexual Behavior: Study 2

Country	Latent mean	Nonadherence	Sample size
Belgium	−0.55	19%	437
Brazil	−0.46	5%	400
Canada	−0.45	13%	334
Denmark	−0.01	7%	394
Estonia	−0.24	23%	255
France	−0.45	27%	408
Germany	−0.48	7%	365
India	−1.23	29%	316
Italy	−0.81	11%	376
Japan	−0.81	5%	347
Netherlands	−0.75	20%	404
Poland	−0.38	9%	393
Portugal	−0.29	21%	319
Singapore	−1.13	10%	320
Switzerland	−0.42	14%	346
United Kingdom	−0.49	15%	348
United States	0.00	12%	433

Note. Countries sorted alphabetically. Parameters are relative to the United States for model identification.

Age differences within countries. Figure 7 displays the means for the different age groups within each country. Countries were sorted in such a way that the means for the youngest age group were in ascending order. In order to have sufficient statistical power, the plot does not present the (small) oldest age groups in Brazil, Estonia, India, Poland, Portugal, and Singapore.

A strong pattern emerged. Unsurprisingly, older participants reported significantly less permissive sexual behavior than younger participants did, with the middle age group somewhere in between. A series of F tests with age as factor and permissive sexual behavior as dependent variable yielded significant differences (i.e., $p < .05$) for all countries, except India, Poland, and Switzerland. In India and Poland, the number of participants in the older age group was very small, which may reduce the power of the test. Marginally significant differences were obtained between the young and middle age group in India but not in Poland.

Cross-country variation. Table 8 also lists the latent means for the countries. Note that the latent mean differences were meaningful, but that the latent scale had no absolute origin. There were significant differences in permissive sexual behavior across countries, $F(16, 6178) = 44.4$, $p < .05$. The US and Denmark reported the highest and Singapore and India reported the lowest prevalence of permissive sexual behavior on the present measure.

Schmitt (2005) found support for the hypothesis that a low availability of men in nations with low sex ratios leads to more promiscuous sex in general. In support of the sex ratio hypothesis, Table 7 shows that a lower sex ratio implied a higher score on permissive sexual behavior ($r = .54$, $p < .05$). The correlation was larger than the one reported by Schmitt. The correlations of permissive sex with low

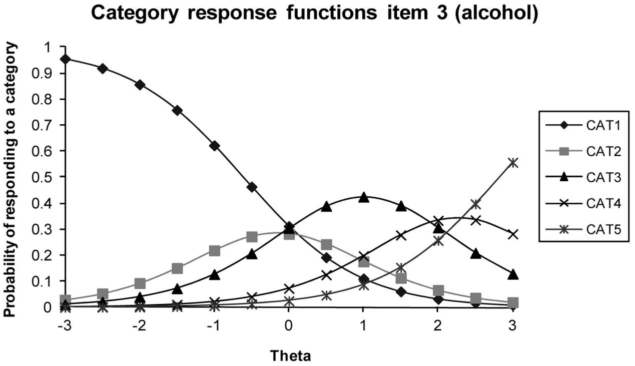

Figure 5. Item characteristic curves, Item 3.

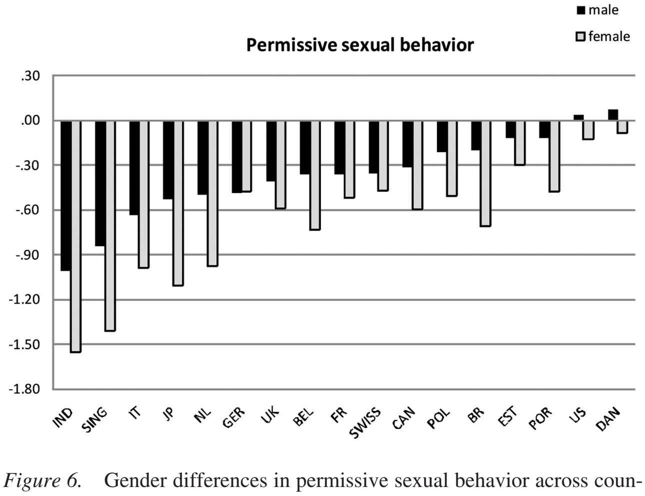

Figure 6. Gender differences in permissive sexual behavior across countries (from left to right, India, Singapore, Italy, Japan, Netherlands, Germany, United Kingdom, Belgium, France, Switzerland, Canada, Poland, Brazil, Estonia, Portugal, United States, Denmark).

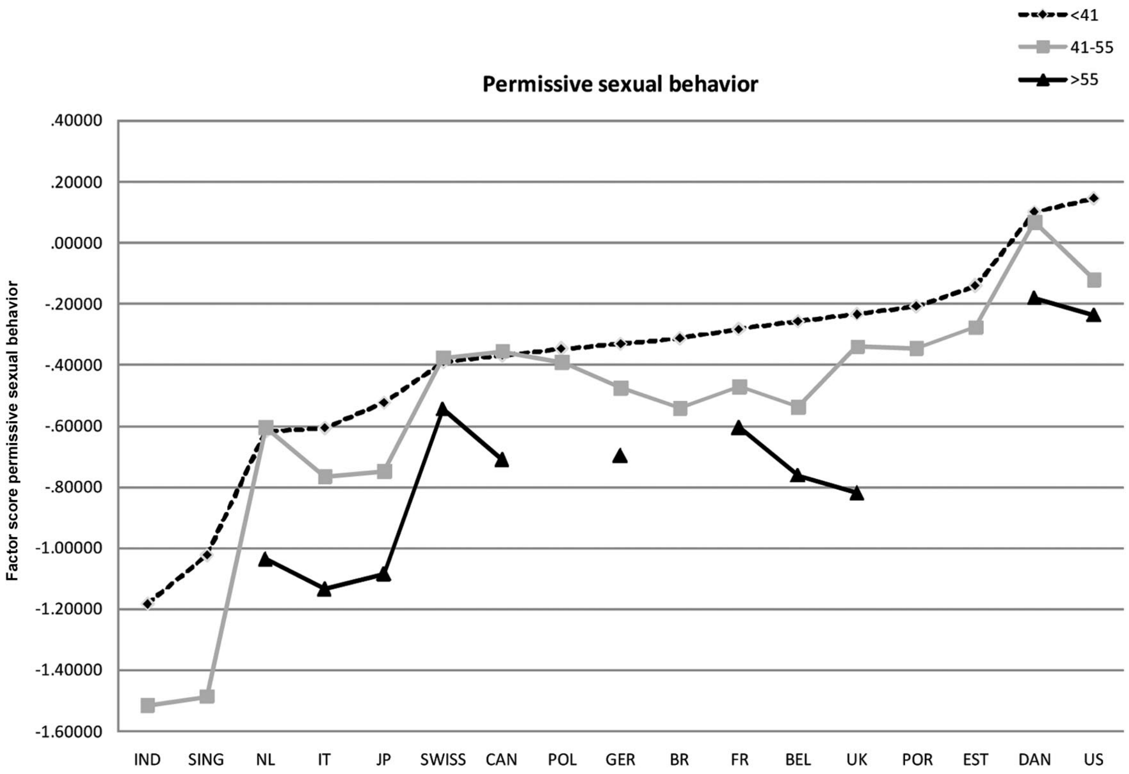

Figure 7. Age differences in permissive sexual behavior across countries (from left to right, India, Singapore, Netherlands, Italy, Japan, Switzerland, Canada, Poland, Germany, Brazil, France, Belgium, United Kingdom, Portugal, Estonia, Denmark, United States).

birth weight infants ($r = -.62$, $p < .01$) and infant mortality ($r = -.46$, $p < .05$) were supportive of strategic pluralism theory (Gangestad & Simpson, 2000). Significant correlations were obtained, again in the same order of magnitude as Schmitt (2005). Thus, Schmitt's findings generalized to nonstudent samples. Finally, the correlation between women's development, especially the percentage of women-headed households, and national levels of permissive sex was remarkably high ($r = .77$, $p < .01$).

Cross-country variation in gender differences. Prior research found that in more demanding reproductive environments, the size of the difference in sociosexuality between men and women was larger (Schmitt, 2005). Table 8 shows that lack of economic resources contributed to the size of the gender gap in permissive sexual behavior, although this factor may be correlated with other sociocultural factors as well. Still, lack of resources appeared associated with women being more careful in partner selection. In particular, permissive sex correlated with GDP/capita ($r = -.44$, $p < .05$), and marginally correlated with the human development index ($r = -.33$, $p < .10$). Another environmental demand variable, the percentage of children with low birth weight, was also marginally significant ($r = .45$, $p < .05$). Furthermore, the level of women's development was also associated with the size of the difference. For instance, the percentage of women-headed households and gender empowerment significantly correlated with the gender gap ($r = -.58$, $p < .05$, $r = -.43$, $p < .05$), whereas the percentage of women-headed households correlated marginally ($r = -.35$, $p < .10$). The signs of these correlations indicated that nations with higher women development (country-level) had smaller gender differences in permissive sexual behavior (person-level).

Variation in gender differences across age groups and countries. The survey design also allowed for an investigation of age-by-gender interactions for permissive sex. Figure 8 presents three plots. The first plot shows the mean permissive sexual behavior for the youngest men and women, and so forth. The size of the bars indicates the size of the differences between men and women, with the shaded areas indicating higher scores for men than for women. Countries were sorted such that the magnitude of the bars increased. That is, India (displayed on the right of the *x*-axis) had the largest reported difference between men and women for the age group < 41 years.

The *F* tests for age-by-gender interactions indicated the presence of significant interactions in a number of countries. There were significant interactions between age and gender for reported permissive sexual behavior in Canada and Germany ($p < .05$). The figure showed that the gender patterns were not always the same across the various age groups. For instance, in Canada and Germany, also in the US, UK, France, Denmark, and Switzerland, there were no significant differences between men and women in the age group below 41.

Discussion of Results

Study 2's rich sampling frame provided several important insights. In line with expectations, the age results in this study

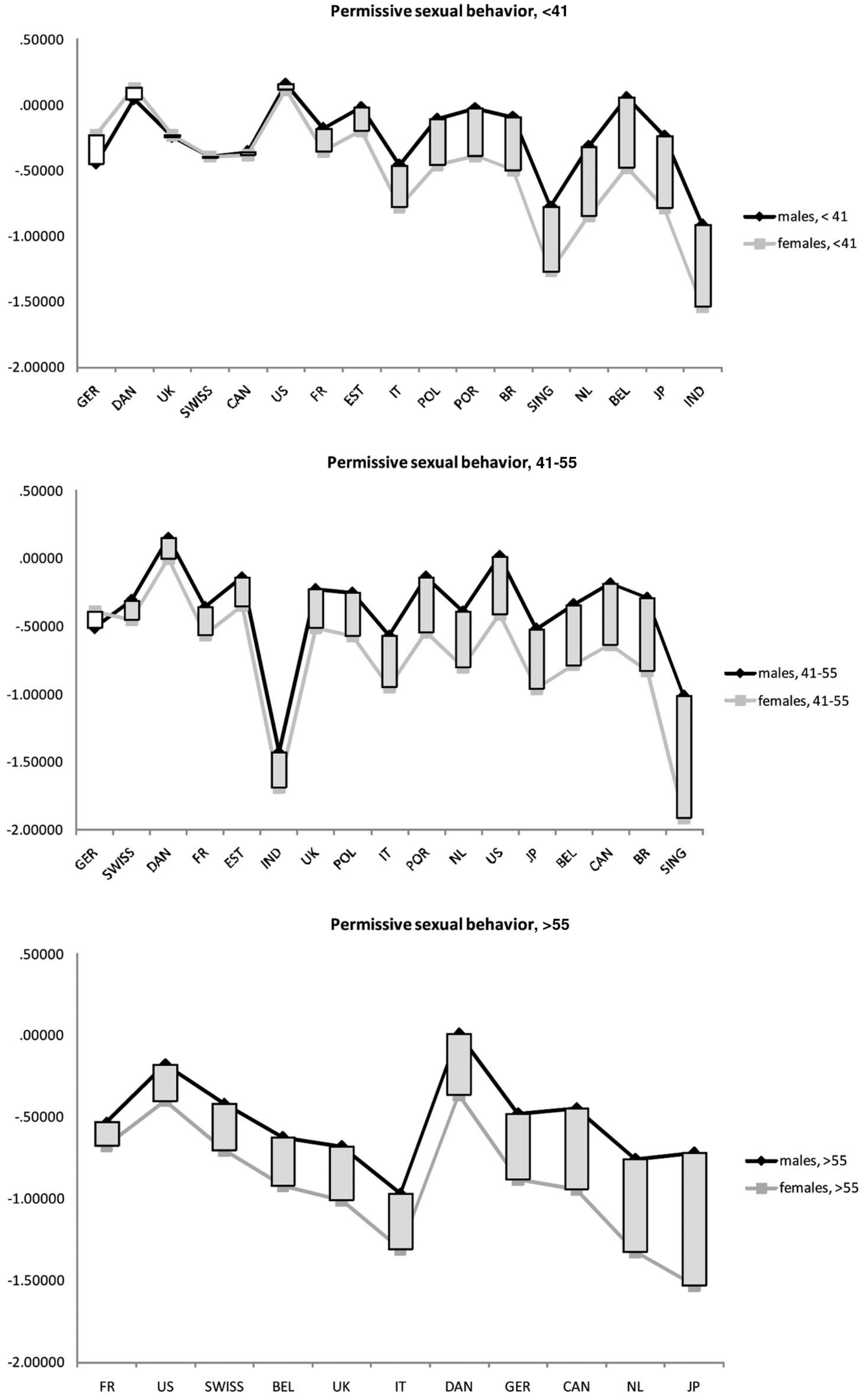

Figure 8. Age-by-gender differences in permissive sexual behavior across countries. GER = Germany; DAN = Denmark; UK = United Kingdom; SWISS = Switzerland; CAN = Canada; US = United States; FR = France; EST = Estonia; IT = Italy; POL = Poland; POR = Portugal; BR = Brazil; SING = Singapore; NL = Netherlands; JP = Japan; Ind = India.

showed that younger age groups displayed more permissive sexual behavior than did older age groups. Furthermore, although men tended to engage in more permissive sexual behavior than women do, this gender difference was not universal. The magnitude of gender differences across age groups and cultures showed there were several predictable conditions under which gender differences wane or even vanish.

National gender differences in permissive sexual behavior varied predictably as a function of sex ratios, familial stress, and women's development. There were no gender differences in permissive sexual behavior for a number of the seventeen countries, and this was especially the case with the younger age group (<41). For this age group, seven countries (41% of the sample countries) did not display significant gender differences, which is in contrast to the findings by Schmitt (2005), who found significant differences in all his sample countries including the US (where the present study found no significant sex differences).

The smaller gender differences among younger than among older people were salient, despite differences across cultures. One possible explanation is that sex differences become larger when people age. However, the finding may also be due to a cohort (time) effect due to changes in division of labor, values, and roles of men and women, which is in line with social structural theory (Eagly et al., 2004; Eagly & Wood, 1999). Indeed, the use of contraceptives in the sampled countries was related to permissive sex differences. A higher use of contraceptives was associated with smaller gender differences in permissive sexual behavior, although causality may be bidirectional. Future research using longitudinal panel data may disentangle such gender, age, cohort, and period effects (Sassler, 2010).

At the country level, permissive sexual behavior in Study 2 correlated predictably with national sex ratio, environmental demand variables, women's development, and use of contraceptives. In line with strategic pluralism strategy, which states that mating strategies depend on local environmental conditions, permissive sexual behavior was lower in demanding environments characterized by higher familial stress. Higher level of women's development and the use of contraceptives were associated with more permissive sexual behavior.

General Discussion

Questions on sensitive topics often make people uneasy about the possible repercussions of disclosing their true answers. Participants are unlikely to answer such questions truthfully in surveys about these topics. The validity of the responses and of the theories and policy measures that are based on these responses is thus threatened. In cross-national research, the issue is exacerbated when the social norms that affect the acceptability of certain responses vary markedly across cultural groups. Thus, any (cross-cultural) inquiry into the psychology of sensitive behavior has to be in tandem with a methodology that addresses this vexing issue of valid reporting. Validity is especially an issue when adult participants are recruited for panel studies of survey and market research organizations, as these collect and connect massive amounts of private and potentially sensitive information. Nevertheless, to improve current theories about fundamental aspects of human behavior, researchers need to go beyond student samples drawn from Western societies (Henrich et al., 2010). Often, the only way to sample such participants is via survey and market research organizations.

The present research proposed MIRRT and applied it to sexual attitudes and behavior. The empirical tests revealed clear cross-cultural differences in sexual activity that were reliably related to characteristics of individuals and nations and had face validity. The new methodology allows for theory building in sensitive domains, such as risky sexual behavior (Catania et al., 1990; Tourangeau & Yan, 2007). The discussion below summarizes the novel aspects of the presented methodology, its implications for theory and practice, and the limitations and opportunities for future research.

Item Randomized Response Theory (IRRT)

Socially desirable responding is problematic. There is a large literature examining the theoretical underpinnings of such response tendencies and their measurement (see Steenkamp, de Jong, & Baumgartner, 2010, for an overview). Although these collective efforts have produced many theoretical insights, they have not yielded an acceptable method to analyze sensitive questions (Tourangeau & Yan, 2007). Univariate randomized response techniques to safeguard anonymous responding have been around for several decades. However, the combination of item response theory (IRT) and randomized response (RR)—resulting in IRRT—is a recent development. It provides new opportunities to gain insight into sensitive issues that are of theoretical and practical concern. The current article extended IRRT and introduced MIRRT, a method that can be applied in large-scale comparative research while accounting for people who do not adhere to the randomized response procedure. The resulting latent scores can then be compared across countries and related to other nomological variables for theory building and testing.

Thus, the advantages of MIRRT are (a) the privacy of participants is protected at the item level, (b) participants are not deceived, (c) over- and underreporting can be controlled at the same time, (d) participants who do not adhere to the procedure can be identified and their effects controlled for, (e) multiple samples can be analyzed simultaneously, and (f) sensitive behavior can be linked to country-level as well as individual-level characteristics. As such, the procedure contributes to the methodological tool kit of (cross-cultural) psychologists and other social science researchers who deal with sensitive questions. To facilitate wider applications of MIRRT, Appendix A contains the code to conduct the analyses in the free program WinBugs.

Theoretical Implications

In Study 1, most of the associations of the Big Five personality dimensions with sexual attitudes and behavior were in line with prior literature, which is reassuring. However, an interesting finding in Study 1 was that extraversion did not relate significantly to risky sexual behavior. If extraverts admit such behavior more easily than introverts when using direct questions, the effect of extraversion found in previous studies might be due to the specific data collection method (a response effect) rather than to intrinsic factors (Schenk & Pfrang, 1986; Schmitt, 2004). Follow-up research may investigate this issue further, for instance, by administering our measure of risky sexual behavior under both direct questioning and randomized response questioning.

Study 2 also produced new insights. The analysis of gender-by-age differences across countries provided evidence for the interactive effects of evolved, dispositional sources and more social structural sources of gender differences in sexual behaviors. The absence of gender differences in several countries and the relationship between sociocultural characteristics at the country level, such as the percentage of women-headed households and the gender gap, were suggestive of more social structural determinants. The general significant gender effects across countries pointed to more evolved, dispositional sources as well. The present findings point to an interactive-source theory of gender effects in sexuality. To disentangle the influence of various sources of gender difference and to estimate their effect sizes, longitudinal research across countries across a sufficiently low period of time is needed. Due to the advent of wide-scale internet panels managed by international survey organizations, cross-national longitudinal studies are no longer difficult or impossible to conduct. Such cross-national web panel surveys make it possible to effectively establish how gender, in interaction with age, period, and cohort effects, determines sexual activity. In combination with MIRRT this facilitates strong tests of the interactive effects of sources of sexuality over the life course.

Practical Implications

Twenty years ago, Catania et al. (1990) called for more research on high-risk sexual behavior, particularly on developing valid methodology for conducting sex research. Despite this call, nearly all sex research to date still resort to direct questioning which places the validity of the findings at risk (Tourangeau & Yan, 2007). The societal importance of staving the spread of sexually transmitted diseases and other undesirable consequences of risky sexual behavior should be sufficient to push researchers to seek out better measurement tools. The new methodologies come at the cost of increased methodological investment. Yet, the IRRT methods developed here are easy to program in public domain software such as WinBugs and require only a few lines of code. Appendix A presents WinBugs code for a two-country study. This source code can easily be extended to accommodate more countries, and it allows researchers to directly implement the method.

IRRT is useful not only for sex research but also for other sensitive domains. Domains that are of substantive interest to psychologists would include, but are not limited to drug use, abortion, alcohol use, mental illnesses, and other conditions involving shame, such as erectile dysfunction, skin diseases, and marital violence. Based on the findings of the present study, it is clear that domain sensitivity interacts with the RR procedure, and it would be interesting to compare nonadherence rates across these various domains.

Limitations and Future Research

A number of caveats are in order. Although this study emphasizes the advantages of randomized response, it is worth noting that, like any technique dealing with social desirability, there are limitations and possibilities for further research. First, MIRRT adds heterogeneity in item responses due to the randomization device. The increased heterogeneity reduces the power of tests, especially if one wishes to test interactions based on small subgroups. Larger sample sizes and some complexity in data analysis are needed to compensate for this, and both are additional costs.

Second, for very sensitive behaviors, relatively high procedural nonadherence rates can be expected. Procedural nonadherence reduces the effective sample size and thereby the power of the tests. Moreover, it is important to know how nonadhering participants actually score on the construct of interest and what the determinants of nonadherence are. Böckenholt and van der Heijden (2007) mention instructional clarity as a potentially important driver of procedural adherence. Indeed, IRRT adds some complexity to the survey response process, and this may introduce random and systematic bias when participants do not adhere to the procedure as a consequence. Another important factor driving nonadherence could be individual and country differences in privacy concerns. Some people may be less willing to share sensitive information than others because they are uncomfortable that professional survey organizations or market research companies will know intimate details of their lives. Perhaps counterintuitively, participants who actually do not engage in the behavior in question but who have to respond *Often* due to the outcome of the randomization device might find it particularly embarrassing to give this forced response and instead do not adhere (Dahl, Manchanda, & Argo, 2001). A fuller understanding of the psychological makeup of nonadhering participants, their motivations, the effect of response options, and how to prevent and accommodate nonadherence are all useful research avenues.

Moving from methodological issues to sampling, it should be noted that although the country sample in Study 2 was large, the sampling frame did not cover some important African and Asian countries due to cost considerations. Including such countries would enable more powerful tests of the dual influence of culture and evolutionary forces on sexual practices. It would also allow for a fuller constellation of sociocultural characteristics that may influence risky sexual behavior. A number of the cultural correlations suggest that a larger sample size may make them significant. Taken together, the present results show the potential of MIRRT and call for more research on adult sexual attitudes and behaviors in different sociocultural environments.

References

Abul-Ela, A. L. A., Greenberg, B. G., & Horvitz, D. G. (1967). A multi proportions randomized response model. *Journal of the American Statistical Association, 62,* 990–1008.

Anderson, J. E., & Dahlberg, L. L. (1992). High-risk sexual behavior in the general population. Results from a national survey, 1988–1990. *Sexually Transmitted Diseases, 19,* 320–325. doi:10.1097/00007435-199211000-00004

Böckenholt, U., & van der Heijden, P. G. M. (2007). Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika, 72,* 245–262. doi:10.1007/s11336-005-1495-y

Boeije, H., & Lensvelt-Mulders, G. (2002). Honest by chance: A qualitative interview study to clarify participants' (non)-compliance with computer-assisted randomized response. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 75,* 24–39. doi: 10.1177/075910630207500104

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review, 100,* 204–232. doi:10.1037/0033-295X.100.2.204

Catania, J. A., Gibson, D. R., Chitwood, D. D., & Coates, T. J. (1990). Methodological problems in AIDS behavioral research: Influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin, 108,* 339–362. doi:10.1037/0033-2909.108.3.339

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95,* 1005–1018. doi:10.1037/a0013193

Clark, R. D., & Hatfield, E. (1989). Gender differences in receptivity to sexual offers. *Journal of Psychology & Human Sexuality, 2,* 39–55. doi:10.1300/J056v02n01_04

Clark, S. J., & Desharnais, R. H. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3,* 160–168. doi:10.1037/1082-989X.3.2.160

Cooper, M. L. (2010). Toward a person × situation model of sexual risk-taking behaviors: Illuminating the conditional effects of traits across sexual situations and relationship contexts. *Journal of Personality and Social Psychology, 98,* 319–341. doi:10.1037/a0017785

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81,* 322–331. doi:10.1037/0022-3514.81.2.322

Dahl, D. W., Manchanda, R. V., & Argo, J. J. (2001). Embarrassment in consumer purchase: The roles of social presence and purchase familiarity. *Journal of Consumer Research, 28,* 473–481. doi:10.1086/323734

David, S., & Knight, B. G. (2008). Stress and coping among gay men: Age and ethnic differences. *Psychology and Aging, 23,* 62–69. doi:10.1037/0882-7974.23.1.62

de Jong, M. G., Pieters, F. G. M., & Fox, J. P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research, 47,* 14–27. doi:10.1509/jmkr.47.1.14

de Jong, M. G., Steenkamp, J. B. E. M., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34,* 260–278. doi:10.1086/518532

Diener, E., & Diener, M. (1995). Cross-cultural correlates of life satisfaction and self-esteem. *Journal of Personality and Social Psychology, 68,* 653–663. doi:10.1037/0022-3514.68.4.653

Eagly, A. H., Diekman, A. B., Johannesen-Schmidt, M. C., & Koenig, A. M. (2004). Gender gaps in sociopolitical attitudes: A social psychological analysis. *Journal of Personality and Social Psychology, 87,* 796–816. doi:10.1037/0022-3514.87.6.796

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior. *American Psychologist, 54,* 408–423. doi:10.1037/0003-066X.54.6.408

Eagly, A. H., & Wood, W. (2005). Universal sex differences across patriarchal cultures ≠ evolved psychological dispositions. *Behavioral and Brain Sciences, 28,* 281–283. doi:10.1017/S0140525X05290052

Edgell, S. E., Himmelfarb, S., & Duncan, K. L. (1982). Validity of forced response in a randomized response model. *Sociological Methods & Research, 11,* 89–100. doi:10.1177/0049124182011001005

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86,* 122–133. doi:10.1037/0021-9010.86.1.122

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Eysenck Personality Questionnaire manual.* San Diego, CA: Educational and Industrial Testing Services.

Fenton, K. A., Johnson, A. M., McManus, S., & Erens, B. (2001). Measuring sexual behaviour: Methodological challenges in survey research. *Sexually Transmitted Infections, 77,* 84–92. doi:10.1136/sti.77.2.84

Fox, J. A., & Tracy, P. E. (1986). *Randomized response: A method for sensitive surveys.* Beverly Hills, CA: Sage.

Fox, J. P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral Statistics, 30,* 189–212. doi:10.3102/10769986030002189

Fox, J. P., & Meijer, R. R. (2008). Using item response theory to obtain individual information from randomized response data: An application using cheating data. *Applied Psychological Measurement, 32,* 595–610. doi:10.1177/0146621607312277

Fox, J. P., & Wyrick, C. (2008). A mixed effects randomized item response model. *Journal of Educational and Behavioral Statistics, 33,* 389–415. doi:10.3102/1076998607306451

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment scales. *Journal of Personality and Social Psychology, 78,* 350–365. doi:10.1037/0022-3514.78.2.350

Gangestad, S. W., & Simpson, J. A. (2000). The evolution of human mating: Trade-offs and strategic pluralism. *Behavioral and Brain Sciences, 23,* 573–587. doi:10.1017/S0140525X0000337X

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis.* Boca Raton, FL: Chapman & Hall.

Guttentag, M., & Secord, P. F. (1983). *Too many women? The sex-ratio question.* Beverly Hills, CA: Sage.

Heaven, P. C. L., Fitzpatrick, J., Craig, F. L., Kelly, P., & Sebar, G. (2000). Five personality factors and sex: Preliminary findings. *Personality and Individual Differences, 28,* 1133–1141. doi:10.1016/S0191-8869(99)00163-4

Hendrick, C., Hendrick, S. S., & Reich, D. A. (2006). The Brief Sexual Attitudes Scale. *Journal of Sex Research, 43,* 76–86. doi:10.1080/00224490609552301

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences, 33,* 61–83. doi:10.1017/S0140525X0999152X

Hogan, J., & Ones, D. S. (1997). Conscientiousness and integrity at work. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 849–870). San Diego, CA: Academic Press.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144. doi:10.1080/03610739208253916

Hoyle, R. H., Fejfar, M. C., & Miller, J. D. (2000). Personality and sexual risk taking: A quantitative review. *Journal of Personality, 68,* 1203–1231. doi:10.1111/1467-6494.00132

Kennedy, G. J., Martinez, M. M., & Garo, N. (2010). Sex and mental health in old age. *Primary Psychiatry, 17,* 22–30.

Kenrick, D. T., Keefe, R. C., Bryan, A., Barr, A., & Brown, S. (1995). Age preferences and mate choice among homosexuals and heterosexuals: A case for modular psychological mechanisms. *Journal of Personality and Social Psychology, 69,* 1166–1172. doi:10.1037/0022-3514.69.6.1166

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology, 90,* 165–178. doi:10.1037/0022-3514.90.1.165

Lalwani, A. K., Shrum, L. J., & Chiu, C. (2009). Motivated response styles: The role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology, 96,* 870–882. doi:10.1037/a0014622

Lameiras Fernández, M., & Rodríguez Castro, Y. (2003). The Big Five and sexual attitudes in Spanish students. *Social Behavior and Personality, 31,* 357–362. doi:10.2224/sbp.2003.31.4.357

Laumann, E. O., Paik, A., & Rosen, R. C. (1999). Sexual dysfunction in the United States: Prevalence and predictors. *JAMA: Journal of American Medical Association, 281,* 537–544. doi:10.1001/jama.281.6.537

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research, 33,* 319–348. doi:10.1177/0049124104268664

Li, N. P., & Kenrick, D. T. (2006). Sex similarities and differences in preferences for short-term mates: What, whether, and why. *Journal of Personality and Social Psychology, 90,* 468–489. doi:10.1037/0022-3514.90.3.468

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Win-

BUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337. doi:10.1023/A:1008929526011

McCrae, R. R., & Costa, P. T. (1997). Conceptions and correlates of openness to experience. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 825–847). San Diego, CA: Academic Press.

McCrae, R. R., & Terracciano, A. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89,* 407–425. doi:10.1037/0022-3514.89.3.407

Moshagen, M., Musch, J., Ostapczuk, M., & Zhao, Z. (2010). Reducing socially desirable responding in epidemiological surveys by a cheating detection extension of the randomized response technique. *Epidemiology, 21,* 379–382. doi:10.1097/EDE.0b013e3181d61dbc

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81,* 660–679. doi:10.1037/0021-9010.81.6.660

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Erlbaum.

Pedersen, F. A. (1991). Secular trends in human sex ratios: Their influence on individual and family behavior. *Human Nature, 2,* 271–291. doi:10.1007/BF02692189

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78,* 582–593. doi:10.1037/0022-3514.78.3.582

Rammstedt, B., & John, O. J. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41,* 203–212. doi:10.1016/j.jrp.2006.02.001

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566. doi:10.1037/0033-2909.114.3.552

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin, 114,* 363–375. doi:10.1037/0033-2909.114.2.363

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Sassler, S. (2010). Partnering across the life course: Sex, relationships, and mate selection. *Journal of Marriage and Family, 72,* 557–575. doi:10.1111/j.1741-3737.2010.00718.x

Schenk, J., & Pfrang, H. (1986). Extraversion, neuroticism, and sexual behavior: Interrelationships in a sample of young men. *Archives of Sexual Behavior, 15,* 449–455. doi:10.1007/BF01542309

Schmitt, D. P. (2003). Universal sex differences in the desire for sexual variety: Tests from 52 nations, 6 continents, and 13 islands. *Journal of Personality and Social Psychology, 85,* 85–104. doi:10.1037/0022-3514.85.1.85

Schmitt, D. P. (2004). The Big Five related to risky sexual behaviour across 10 world regions: Differential personality associations of sexual promiscuity and relationship infidelity. *European Journal of Personality, 18,* 301–319. doi:10.1002/per.520

Schmitt, D. P. (2005). Sociosexuality from Argentina to Zimbabwe: A 48-nation study of sex, culture, and strategies of human mating. *Behavioral and Brain Sciences, 28,* 247–275. doi:10.1017/S0140525X05000051

Schwartz, S. H., & Rubel-Lifschitz, T. (2009). Cross-national variation in the size of sex differences in values: Effects of gender equality. *Journal of Personality and Social Psychology, 97,* 171–185. doi:10.1037/a0015546

Segura, S. L., & González-Romá, V. (2003). How do participants construe ambiguous response formats of affect items? *Journal of Personality and Social Psychology, 85,* 956–968. doi:10.1037/0022-3514.85.5.956

Simpson, J. A. (2009). Editorial. *Journal of Personality and Social Psychology, 96,* 60. doi:10.1037/a0014798

Simpson, J. A., & Gangestad, S. W. (1991). Individual differences in sociosexuality: Evidence for convergent and discriminant validity. *Journal of Personality and Social Psychology, 60,* 870–883. doi:10.1037/0022-3514.60.6.870

Simpson, J. A., & Gangestad, S. W. (1992). Sociosexuality and romantic partner choice. *Journal of Personality, 60,* 31–51. doi:10.1111/j.1467-6494.1992.tb00264.x

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298–321. doi:10.1177/0146621605285517

Smith, D. B., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87,* 211–219. doi:10.1037/0021-9010.87.2.211

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire stress reaction scale. *Journal of Personality and Social Psychology, 75,* 1350–1362. doi:10.1037/0022-3514.75.5.1350

Steenkamp, J. B. E. M., de Jong, M. G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research, 47,* 199–214. doi:10.1509/jmkr.47.2.199

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133,* 859–883. doi:10.1037/0033-2909.133.5.859

Trobst, K. K., Wiggins, J. S., Costa, P. T., Jr., Herbst, J. H., McCrae, R. R., & Masters, H. L., III. (2000). Personality psychology and problem behaviors: HIV risk and the five-factor model. *Journal of Personality, 68,* 1233–1252. doi:10.1111/1467-6494.00133

Tuerlinckx, F., De Boeck, P., & Lens, W. (2002). Measuring needs with the Thematic Apperception Test: A psychometric study. *Journal of Personality and Social Psychology, 82,* 448–461. doi:10.1037/0022-3514.82.3.448

Turchik, J. A., & Garske, J. P. (2009). Measurement of sexual risk taking among college students. *Archives of Sexual Behavior, 38,* 936–948. doi:10.1007/s10508-008-9388-z

United Nations Development Programme. (2001). *Human development report 2001.* New York, NY: Oxford University Press.

United Nations Statistics Division. (2001). *World population prospects: The 2000 revision. Vol. 1: Comprehensive tables.* New York, NY: United Nations.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60,* 63–69.

Weinhardt, L. S., Forsyth, A. D., Carey, M. P., Jaworski, B. C., & Durant, L. E. (1998). Reliability and validity of self-report measures of HIV-related sexual behavior: Progress since 1990 and recommendations for research and practice. *Archives of Sexual Behavior, 27,* 155–180. doi:10.1023/A:1018682530519

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89,* 696–716. doi:10.1037/0022-3514.89.5.696

Appendix A

WinBugs Code for Model M_{10}

The WinBugs code below assumes there are two groups.

```
####################################################
#### Beginning of Model Command File
#### Notation
#### a[k] = discrimination parameter of item k
#### beta[k,c] = threshold parameter for category c of item k
#### gmem[i] = non-adherence class membership of respondent i
#### theta[i] = latent construct score of respondent i
#### K = number of items
#### C = number of response categories of each item
#### p1 = probability of having to give an your own answer
#### p2[c] = probability of forced response for category c (probability may vary across categories)
#### mu2 = mean of latent construct in group 2
#### va2 = precision of latent construct in group 2
#### p_NA1 = probability of belonging to non-adherence class in group 1
#### p_NA2 = probability of belonging to non-adherence class in group 2
model {
#### specification of model for group 1. In the example, we assume there are 400 respondents and 5 response options for each item (which implies 5 − 1 = 4 logit expressions in WinBugs).
for (i in 1: 400) {
for (k in 1: K) {
for (c in 1: 4) {
logit(Q[i, k, c]) <- a[k]*(beta[k,c] − theta[i])
}
# specification of category response probability for group 1
p[i,k,1]<-gmem[i]+(1-gmem[i])*(p1*Q[i,k,1]+(1-p1)*p2[1])
p[i,k,2]<-(1-gmem[i])*(p1*(Q[i,k,2]-Q[i,k,1])+(1-p1)*p2[2])
p[i,k,3]<-(1-gmem[i])*(p1*(Q[i,k,3]-Q[i,k,2])+(1-p1)*p2[3])
p[i,k,4]<-(1-gmem[i])*(p1*(Q[i,k,4]-Q[i,k,3])+(1-p1)*p2[4])
p[i,k,5]<-(1-gmem[i])*(p1*(1-Q[i,k,4])+(1-p1)*p2[5])
Y[i,k]~dcat(p[i,k,])
}
theta[i] ~ dnorm(0,1) # prior for latent construct score in group 1 (baseline group)
gmem[i] ~ dbern(p_NA1) # prior for non-adherence class membership in group 1
}
#### specification of model for group 2. Note, it is assumed there are 400 respondents in group 2 and again, 5 response options for each item.
for (i in 401: 800) {
for (k in 1: K) {
for (c in 1: 4) {
logit(Q[i, k, c]) <- a[k]*(beta[k,c] − theta[i])
}
# specification of category response probability for group 2
p[i,k,1]<-gmem[i]+(1-gmem[i])*(p1*Q[i,k,1]+(1-p1)*p2[1])
p[i,k,2]<-(1-gmem[i])*(p1*(Q[i,k,2]-Q[i,k,1])+(1-p1)*p2[2])
p[i,k,3]<-(1-gmem[i])*(p1*(Q[i,k,3]-Q[i,k,2])+(1-p1)*p2[3])
p[i,k,4]<-(1-gmem[i])*(p1*(Q[i,k,4]-Q[i,k,3])+(1-p1)*p2[4])
p[i,k,5]<-(1-gmem[i])*(p1*(1-Q[i,k,4])+(1-p1)*p2[5])
Y[i,k]~dcat(p[i,k,])
}
theta[i]~dnorm(mu2,va2) # prior for latent construct score in group 2
gmem[i] ~ dbern(p_NA2) # prior for non-adherence class membership in group 2
}
# specification of priors for item parameters
for (k in 1: K) {
# log-normal prior for discrimination parameter of item k
a[k]~dlnorm(0,0.5)
# prior for thresholds based on truncated normal
# if numerically unstable, impose normal prior on log threshold differences
# prior for threshold 1 of item k
beta[k,1] ~ dnorm(0,0.1)I(, beta[k,2])
# prior for threshold 2 of item k
beta[k,2] ~ dnorm(0,0.1)I(beta[k,1], beta[k,3])
# prior for threshold 3 of item k
beta[k,3] ~ dnorm(0,0.1)I(beta[k,2], beta[k,4])
# prior for threshold 4 of item k
beta[k,4] ~ dnorm(0,0.1)I(beta[k,3],)
}
# specification of priors for hyperparameters
# beta prior for non-adherence probability in group 1
p_NA1 ~ dbeta(1,1)
# beta prior for non-adherence probability in group 2
p_NA2 ~ dbeta(1,1)
# normal prior for latent construct mean in group 2
mu2 ~ dnorm(0,0.1)
# gamma prior for latent construct precision in group 2
va2 ~ dgamma(1,1)
}
#### end of model command file
#### Beginning of Data List
```

(*Appendices continue*)

```
#### In the example, the randomized response probabilities are p1 = 2/3, p2[1]= p2[2]= p2[3]= p2[4]=1/6, #### and p2[5]=2/6, and there are 10 items. The total number of respondents is 400 + 400 = 800
List(K = 10, p1 = 0.666, p2 = c(0.1666, 0.1666, 0.1666, 0.1666, 0.3666),
Y = structure(.Data = c(1, 3, 4, 5, 5, 4, 2, 4, 5, 2,
. . .
. . .
. . ., .Dim = c(800,10)))
#### End of Data List
#### Beginning of Initial Values List
####
#### No initial values are specified for the threshold parameters, but there are initial values for the
#### discrimination parameters
List(mu2 = 0, va2 = 1, p_NA1 = 0.1, p_NA2 = 0.1, a=(0,0,0,0,0,0,0,0,0,0))
#### End of Initital Values List
###########################################################
```

Appendix B

Measurement Invariance

This appendix describes the steps to test for measurement invariance if there are two countries. The analysis starts with a baseline model (M_0) in which item parameters are allowed to vary across the two countries. Although this model is identified (if the mean and variance in one of the countries is fixed to 0 and 1), the scale is not on a common metric across countries. However, the fit of this model can be compared to the fit of nested models with more restrictions.

One or more anchor items are designated to establish a common metric (Reise, Widaman, & Puch, 1993). An anchor item has invariant item parameters across groups. A model with full measurement invariance imposes similar item parameters for all items, and because the item parameters do not vary across countries, the metric is common (M_1). If this model fits worse than the baseline model (M_0), a model with partial measurement invariance is used (M_2).

Model selection (M_0 to M_2) can be established via a likelihood ratio statistic of nested models. Even though estimation is in a Bayesian framework, the parameter estimates are virtually identical to those obtained via maximum likelihood when uninformative priors are used, in which case the likelihood ratio statistic from frequentist statistics provides a useful diagnostic check. Alternatively, the deviance information criterion (DIC) statistic or Bayes factors can be used to compare different nested models. In Study 1, the various procedures lead to the same conclusions.

In WinBugs, the log-likelihood of a model can be obtained via the output that is produced for the DIC statistic. The Dbar statistic is equal to −2*log-likelihood of a given model, and hence a simple calculation produces the estimated log-likelihood that can be compared to the log-likelihood of a nested model. The likelihood-ratio statistic has a chi-squared distribution with degrees of freedom equal to the number of test parameters. Models with partial measurement invariance can be specified on an item-by-item basis. For one item at a time, the item parameters are constrained and the likelihood ratio test is used to test the plausibility of the restriction.

Appendix C

Bayesian *F* test

For latent means, we can consider $J - 1$ linear contrasts $\Delta_j = \bar{\theta}^j - \bar{\theta}^J$, where $\bar{\theta}^j$ is the mean of θ in country j. Then, $p(\Delta|\mathbf{x})$ is a monotonic decreasing function of a function Q_0 which is asymptotically distributed as $F_{(J\text{-}1,\ N\text{-}J)}$ as $N_j \to \infty$ For large samples, the vector $\Delta_0 = \mathbf{0}$ is included in the highest posterior density (HPD) region of $1 - \alpha$ if and only if:

$$\lim_{N_j \to \infty} P[p(\Delta) > p(\Delta_0)] = P\left(F_{(J-1,\, N-J)} < \frac{\sum_g N_j(\bar{\theta}^j - \bar{\bar{\theta}})^2}{(J-1)s^2} \right) < 1 - \alpha$$

where $\bar{\bar{\theta}} = \frac{1}{N} \sum_j N_j \bar{\theta}^j$ is the overall mean. The hypothesis of equal means across countries is rejected when

$$P\left(F_{(J-1,\, N-J)} < \frac{\sum_j N_j(\bar{\theta}^j - \bar{\bar{\theta}})^2}{(J-1)s^2} \right) > 1 - \alpha.$$

Received December 20, 2010
Revision received April 2, 2012
Accepted June 20, 2012 ■