11-2017

On Enhancing the Cross-Cultural Comparability of Likert-Scale Personality and Value Measures: A Comparison of Common Procedures

Jia He

Fons J.R. Van de Vijver

Velichko H. Fetvadjiev

Alejandra de Carmen Dominguez Espinosa

Byron Adams

See next page for additional authors

Authors
Jia He, Fons J.R. Van de Vijver, Velichko H. Fetvadjiev, Alejandra de Carmen Dominguez Espinosa, Byron Adams, Itziar Alonso-Arbiol, Arzu Aydinli-Karakulak, Carmen Buzea, Radosveta Dimitrova, Alvaro Fortin, Given Hapunda, Sang Ma, Ruta Sargautyte, Samantha Sim, Maja K. Schachner, Angela Suryani, Pia Zeinoun, and Rui Zhang

On Enhancing the Cross-Cultural Comparability of Likert-Scale Personality and Value Measures: A Comparison of Common Procedures

He, J., Van de Vijver, F. J. R., Fetvadjiev, V. H., Dominguez-Espinosa, A., Adams, B. G., Alonso-Arbiol, I., Aydinli-Karakulak, A., Buzea, C., Dimitrova, R., Fortin Morales, A., Hapunda, G., Ma, S., Sargautyte, R., Schachner, R. K., Sim, S., Suryani, A., Zeinoun, P., & Zhang, R. (in press). On enhancing the cross-cultural comparability of Likert-scale personality and value measures: A comparison of common procedures. *European Journal of Personality*.

Abstract

This study aims to evaluate a number of procedures that have been proposed to enhance cross-cultural comparability of personality and value data. A priori procedures (anchoring vignettes and direct measures of response styles (i.e., acquiescence, extremity, midpoint responding, and social desirability), a posteriori procedures focusing on data transformations prior to analysis (ipsatization and item parceling), and two data modeling procedures (treating data as continuous vs. as ordered categories) were compared using data collected from university students in 16 countries. We found that (1) anchoring vignettes showed lack of invariance, so they were not bias-free; (2) anchoring vignettes showed higher internal consistencies than raw scores where all other correction procedures, notably ipsatization, showed lower internal consistencies; (3) in measurement invariance testing, no procedure yielded scalar invariance; anchoring vignettes and item parceling slightly improved comparability, response style correction did not affect it, and ipsatization resulted in lower comparability; (4) treating Likert-scale data as categorical resulted in higher levels of comparability; (5) factor scores of scales extracted from different procedures showed similar correlational patterning; and (6) response style correction was the only procedure that suggested improvement in external validity of country-level conscientiousness. We conclude that, although no procedure resolves all comparability issues, anchoring vignettes, parceling, and treating data as ordered categories seem promising to alleviate incomparability. We advise caution in uncritically applying any of these procedures.

Keywords: personality, values, anchoring vignettes, response styles, score standardization, parceling

On Enhancing the Cross-Cultural Comparability of Likert-Scale Personality and Value Measures: A Comparison of Common Procedures

In cross-cultural research, it is imperative to demonstrate desirable psychometric properties of target measures, in particular their cross-cultural comparability, before any comparative inferences are drawn (van de Vijver & Leung, 2001). Inadequate comparability often proves to be the case in large-scale studies involving multiple cultures. For instance, in their study of personality traits across 56 cultures, Schmitt et al. (2007) reported a comparable structure of the Big Five model, but higher levels of measurement equivalence, such as comparability of scores, were not assessed. Research on basic human values has similarly reported mostly structural, but not metric equivalence (Davidov, Schmidt, & Schwartz, 2008; Schwartz & Sagiv, 1995). With multigroup confirmatory factor analysis as a rigorous means to test scale comparability, studies involving many cultures tend to show levels of comparability that do not allow for direct comparison of means across cultures (e.g., Organisation for Economic Cooperation and Development [OECD], 2014), although such comparisons are often seen as an important reason for their existence.

Different procedures have been proposed to deal with comparability problems. Some involve changes in the design and instruments employed in a study, often aimed at reducing the impact of response styles such as acquiescent response style (ARS, the tendency to always agree), extreme response style (ERS, the tendency to use the end points of a response scale), midpoint response style (MRS, the tendency to use the midpoint of a response scale), and socially desirable responding (SDR, the tendency to respond in a way so as to make oneself look good). Other procedures involve statistical modeling, often aimed at demonstrating equivalence of constructs and scales across cultures. However, the design and analysis approaches are not always well linked; for example, it is rather uncommon to examine to what extent score corrections for response styles, conducted to improve the validity of cross-

cultural comparisons by presumably removing response styles, actually improve the validity of cross-cultural comparisons. In the present study we examine the most common ways of addressing incomparability (outlined in more detail below) and compare their psychometric merits. In the following, we first review the sources of incomparability and the corresponding levels of comparability in cross-cultural research; we then review the main findings on comparability in personality and value research; finally, we describe the main procedures aimed at both design and statistical analysis stages to address the incomparability issue.

Bias and Equivalence

Incomparability stems from bias, which refers to systematic errors that threaten the validity of measures administered in different cultures (van de Vijver & Leung, 1997). The existence of bias implies that differences in observed scores may not correspond to genuine differences in the target construct. Three sources of bias can be distinguished. *Construct bias* means that the target construct has a different meaning across cultures. *Method bias* includes incomparability due to sampling, instruments, and administration. One main source of method bias targeted in this study is scale usage preferences, which are difficult to avoid in Likert-scales. These scale usage preferences can refer to response styles, defined as the systematic tendency to respond to questionnaire items on some basis other than the target construct (Paulhus, 1991) and to reference-group effects resulting from different standards (i.e., a reference group) that respondents use to evaluate themselves and their behaviors (van de Gaer, Grisay, Schulz, & Gebhardt, 2012). A classic example of this method bias is the paradoxical correlations of students' achievement and self-reported motivation in the Programme for International Student Assessment (PISA). In all participating countries, students' self-reported learning motivation is positively related to their achievement in science, mathematics, and reading. However, when scores are aggregated at country level and the correlation is computed between countries' average levels of motivation and achievement, a negative

correlation is found. That is, East Asian countries such as China, Korea, and Japan, typically showing *high* scores on (externally assessed) achievement in the PISA studies, tend to have *low* scores on (self-reported) learning motivation. Thus, the comparability and validity of self-report measures across cultures are challenged. Finally, *item bias* indicates a different probability of endorsing an item given the same trait level of individuals from different cultures.

The presence of bias can threaten the psychological equivalence, that is, the level of comparability of scores across cultures. Three main levels of equivalence (called invariance in the structural equation modeling literature)[1] can be distinguished and statistically tested: (1) Configural invariance means that across cultures the construct is understood in the same way. In statistical terms, this level of invariance signals that items in a measure exhibit the same configuration of salient and non-salient factor loadings. (2) Metric invariance indicates that items of the construct have the same factor loadings across cultures. With metric invariance, scale score comparisons can be made within cultures (e.g., the personality trait of agreeableness can be compared between males and females within each culture), and the association of variables can be compared across cultures (e.g., correlations between agreeableness and openness can be compared across cultures, if both scales reach metric invariance). (3) Scalar invariance implies that items have the same intercepts (i.e., point of origin) across cultures. By implication, scores on scalar-invariant measures have the same psychological meaning across cultures; for example, two individuals from two different cultures who both have a mean score of 4 on an Extraversion scale, can be taken to be equally extraverted. Only with scalar invariance can mean scores of scales be validly compared across cultures (van de Vijver & Leung, 1997).

[1] We use *bias* and *equivalence* as the generic terms for lack of comparability and comparability, respectively, and we use *measurement invariance* as the psychometric term for comparability testing.

Cross-Cultural Comparability in Personality and Values

Cross-cultural research on personality has found substantial evidence for configural invariance, especially of Big Five instruments. Both in the Big Five Inventory (Schmitt et al., 2007) and the Revised NEO Personality Inventory (NEO-PI-R; McCrae, Terracciano, & 78 Members of the Personality Profiles of Cultures Project, 2005), configural invariance has been established across a large number of countries from around the world, albeit with some individual deviations. Measures of other broad personality models, such as Eysenck's Psychoticism–Extroversion–Neuroticism (Barrett, Petrides, Eysenck, & Eysenck, 1998) and Zuckerman's Alternative Five Factor Model (Rossier et al., 2016), have also shown structural equivalence across many countries. Stricter levels of invariance have been examined less often and have proven much more difficult to meet. Nye, Roberts, Saucier, and Zhou (2008) found support only for configural invariance of the Big Five Mini Markers (Saucier, 1994) in three countries. Thalmayer and Saucier (2014) found some support, although weak by conventional standards, for configural invariance of a version the Questionnaire Big Six in 26 countries, but not for Big Five or Big Two versions of the questionnaire, nor for any stricter levels of equivalence. Rossier et al. (2016) found evidence for metric equivalence of the individual scales of the Alternative Five Factor Model in 23 cultures, but much less evidence when the model was tested as a whole. Finally, Church et al. (2011) found evidence for item bias in 40% to 50% of the items of the NEO-PI-R across the USA, Mexico, and the Philippines.

Similarly to personality research, studies of values have mostly found support for configural invariance across cultures (Fischer, 2014). The prevalent technique since the formulation of Schwartz's (1992) circular theory of values has been multidimensional scaling. In this approach, judgments of structure replication are typically based on a visual inspection of the ordering of values on spatial plots (e.g., Bilsky, Janik, & Schwartz, 2011), although

quantitative assessment of replicability is also possible (Fontaine, Poortinga, Delbeke, & Schwartz, 2008). Confirmatory factor analysis (CFA) has provided support for the theoretical structure of the basic values as a common structure for 27 countries (Schwartz & Boehnke, 2004). In one of the rare studies to assess equivalence in CFA, Davidov et al. (2008) found support for metric equivalence, although only after merging several values and allowing cross-loadings of several items onto opposite values.

The limitations of the CFA framework in the context of both personality and values have long been recognized, such as the complexity of these multidimensional models and issues related to the application and interpretation of CFA models (Fischer, 2014; Hopwood & Donnellan, 2010). However, despite the salient problems of cross-cultural comparability of personality and value measures, no previous research has systematically examined the full suite of methods involving design-based and statistical procedures to enhance comparability. Moreover, only a few previous studies involving over a dozen cultures systematically checked comparability within a multigroup confirmatory factor analysis framework. The present study aims to fill this gap.

Procedures for Addressing Equivalence

Both a priori and a posteriori procedures for the detection and correction of bias have been proposed. In this study, we focus on design-based and statistical approaches in personality and value measures. Specifically, we used adapted designs aimed to enhance score comparability (i.e., assessing anchoring vignettes and response styles), statistical correction procedures (i.e., within-subject standardization and parceling), and two statistical modeling procedures (i.e., treating Likert-scale responses as continuous and as categorical) in the evaluation of their effects on data comparability.

Design-Based Procedures for Response Scale Usage Correction

Anchoring vignettes. Anchoring vignettes try to establish a common reference point for respondents from various groups (King, Murray, Salomon, & Tandon, 2004). Respondents are asked to rate several vignettes, which are descriptions of hypothetical persons with different levels of the target trait. Respondents are also asked for a self-assessment using the same response format. Whereas the systematic differences in responses to the same vignette rating supposedly reflect mainly differences in scale usages, responses on self-assessment are a combination of such distortion and the true trait level. Therefore, the measurement bias due to reference group differences from the self-assessment can be removed to yield an estimate of the true trait level.

A prominent example of the use of vignettes in cross-cultural personality research can be found in Mõttus et al. (2012a). These authors used vignettes to examine in 21 countries the so-called conscientiousness paradox, whereby the country-level scores on conscientiousness do not correspond to widely shared stereotypes, and have counterintuitive correlations with external criteria such as gross domestic product per capita (GDP) and longevity. When the original responses in this study were adjusted by means of the vignettes, the country orderings changed little and the correlations with external criteria moved generally in the expected direction, but remained weak. A somewhat larger shift of country-level scores in the expected direction was achieved when vignette scores were used to derive response-style indices (Mõttus et al., 2012b) and raw responses were subsequently corrected by means of these indices. Given the relevance of conscientiousness for some of the most counterintuitive findings in country-level personality scores, the present study also examines reference-group effects in trait conscientiousness, but further explores reference-group effects in vignettes of values and response styles (as described in the Measures section) and examines these effects for the scores on all of the Big Five factors and Schwartz's values (Schwartz & Sagiv, 1995).

There are two working assumptions in the vignette approach: response consistency (i.e., respondents use the same mechanisms to give responses to self-assessment questions and the vignette questions) and vignette equivalence (i.e., the vignettes are understood by all respondents in the same way). The implication is that responses to various vignettes are affected by a common nuisance factor that varies across cultures, and that a score correction based on the raw vignette score in each culture, reduces the impact of this nuisance factor. Existing literature has shown mixed results on the soundness of these assumptions among different populations and topics of interest (e.g., Jürges & Winter, 2013; Kapteyn, Smith, Van Soest, & Vonkova, 2011). Particularly for vignette equivalence as the "golden standard" to applying vignette ratings to rescale self-assessment, its tenability is of utmost importance. This study makes use of multiple sets of vignettes on constructs related to response tendencies, personality and values, which can be used to empirically test the tenability of vignette equivalence.

Response style correction. The response styles that are most frequently studied include ARS, ERS, MRS, and SDR. These four response styles are interrelated; a general response style (GRS) factor can be extracted with positive loadings of ERS and SDR and negative loadings of ARS and MRS, representing response amplification to moderation (He & van de Vijver, 2013). Correction for GRS was found to have limited effects on sizes of cross-cultural differences and country rankings in teachers' self-report data from the Teaching and Learning International Survey (TALIS; He & van de Vijver, 2015a), yet empirical evidence is lacking as to the magnitude of the impact correction effects have on the measurement comparability of data.

The validity of response bias indicators and their use for correcting scores is a continually contested topic in psychological assessment (McGrath, Mitchell, Kim, & Hough, 2010; Rohling et al., 2010). These response styles may be a nuisance that needs to be removed,

but they may also represent valid cultural and individual differences in personality and values (e.g., Smith, 2004, 2011). SDR, perhaps the style that has received most attention in personality psychology, has been found to contain a substantive component (Bäckström & Björklund, 2014; McCrae & Costa, 1983). Furthermore, SDR displays meaningful variation and correlates across cultures (Bou Malham, & Saucier, 2016). Still, no previous research has addressed the effects of response style correction using direct assessment of different response styles on the cross-cultural equivalence of personality or value scores. This is an important omission addressed in the present study.

Statistical Procedures within Target Constructs

Ipsatization. Score standardization involves adjustment procedures using the mean and (less frequently) standard deviation of each respondent (i.e., within-subject) or of each cultural group (i.e., within-group) to control for scale usage differences (Fischer, 2004). Ipsatization (within-subject standardization where the mean of a range of items of an individual is subtracted from the raw response to each item) has often been used in psycholexical studies to obtain clearer factor structures (De Raad et al., 2014). The approach has wider applications, too. Using the balanced 10-item Big Five Inventory, Rammstedt, Kemper, and Borg (2013) found that partialling out the mean rating in personality items (as a way to control for acquiescence) resulted in higher structural equivalence and better correspondence to an ideal Big Five pattern of loadings in 18 cultures. In a study on the effects of cross-role consistency on psychological adjustment, Church et al. (2008) found that ipsatized data were more in line with the theoretical expectations, showing beneficial effects of consistency across six cultures; thus, ipsatization was seen as a successful control for response styles. In value research with the Schwartz Value Survey (SVS) and the Portrait Value Questionnaire (PVQ), ipsatization is also common practice. The underlying theoretical

assumption is that differences in individual means across all value items reflect scale usage and not value substance (Schwartz et al., 2001).

However, ipsatization has several potential problems. Subtracting the individual mean from a set of items may remove genuine, meaningful individual differences when it is unrealistic to assume that the sum of all scores across individuals is identical. As the sum of all ipsatized scores for each respondent is zero, each ipsatized item score is dependent on scores of other items; as a consequence, the average inter-item correlation among these scores is slightly negative. Given the widespread use of ipsatization but also the salient statistical complications associated with it (Fischer, 2004), it is important to compare the merits of this procedure to alternative procedures for increasing cross-cultural equivalence.

Parceling. Parceling is a procedure to combine individual items in a single (parcel) score and use these combined items as indicators of a latent factor. Parcels are advocated based on the idea that such item sets will show less item bias (e.g., when parcels are formed, bias in different items to some extent cancels each other out), and that parcels comply more with distributional assumptions of factor analytic models than do individual items, thus they are more appropriate for maximum likelihood estimation. With unidimensional constructs (i.e., one-factor models), parceling seems to be able to improve model fit and reduce bias in factor analyses (e.g., Bandalos, 2002). However, some studies suggested detrimental effects of using item parcels on measurement invariance tests, especially in multidimensional constructs, where misspecifications of parcels and sources of bias could be camouflaged (Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013). Parceling is thus not common in research on broad and complex personality structure models such as the Big Five. The procedure has been used with more success in research on social-cognitive constructs such as social axioms (Leung et al., 2012), but also on values (Leung et al., 2007). The present study is the first to test the

effectiveness of parceling in enhancing the cross-cultural comparability of personality next to values.

MGCFA Modeling Procedures

Multigroup confirmatory factor analysis (MGCFA) is by far the most rigorous approach for measurement invariance tests (e.g., Byrne, 2001; Cheung & Rensvold, 2002; Rutkowski & Svetina, 2014, 2016). Treating Likert-scale responses as continuous in MGCFA is common practice. In this approach, factor loadings and item intercepts are estimated, and equality constraints are imposed with maximum likelihood estimation (ML) as one of different available estimators. However, common violations of the assumption of multivariate normality in empirical data worsen model fit. Therefore, it has been proposed that it is more realistic to model Likert-scale responses as ordered categories (Rutkowski & Svetina, 2014). In the categorical MGCFA model, parameter estimation and equality constraints apply to factor loadings and item thresholds (number of thresholds equals number of categories minus one), so there is more flexibility. In these models, weighted least squares with mean and variance adjustment (WLSMV) estimation is often used, because this procedure yields a robust estimator that does not assume normally distributed variables and provides the best option for modeling categorical or ordered data (Brown, 2006). It is expected that modeling data as ordered categories would result in higher levels of comparability than treating data as continuous. Treating Likert-scale responses as categorical is a relatively new approach, and has not been used much in cross-cultural research on personality. In the only study we could find, Eigenhuis, Kamphuis, and Noordhof (2015) analyzed the equivalence of the Multidimensional Personality Questionnaire in the Netherlands and the USA and found that about 40% of the items were biased. This percentage was close to results found by Church et al. (2011) using continuous modeling; yet, no previous study has directly compared results

using both methods in personality research. We compare the results of both approaches in the present study.

Model fit in MGCFA can be evaluated by various measures, such as chi-square statistics (though the statistic is very sensitive to sample sizes), Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA). The acceptance of a more restrictive model is based on the change of CFI and RMSEA. In the contexts of large-scale assessment with up to 20 cultures, Rutkowski and Svetina (2014) proposed to set the cut point of ΔCFI to .02 and that of ΔRMSEA to .03 from configural to metric models, and to .01 from metric to scalar models for both ΔCFI and ΔRMSEA when data are treated as continuous (using the ML estimator). For categorical models, Rutkowski and Svetina (2016) proposed ΔCFI of .004 and ΔRMSEA of .05 from configural to metric models, and ΔCFI of .004 and ΔRMSEA of .01 from metric to scalar models (using the WLSMV estimator). To ensure valid conclusions, we follow the most up-to-date recommended model fit criteria in MGCFA for over a dozen countries as specified in Rutkowski and Svetina (2014, 2016).

The Present Study

As our review suggests, different procedures have been designed to enhance the cross-cultural comparability of Likert-scale personality and values data. However, no study to date has addressed the relative merits and limits of each procedure in comparison. The present study aims to address this knowledge gap. On the basis of the existing literature on the replicability of the structure of personality and values across cultures, we do not expect to find evidence of widespread construct bias. However, method bias, expressed in scale usage differences, and item bias are more likely to affect cross-cultural comparisons in these domains. A priori (design) procedures including anchoring vignettes and measures of response style, a posteriori (computational) procedures focusing on data transformations prior to analysis (ipsatization and item parceling), and two data modeling procedures (treating data

as continuous and as ordered categories) are compared. The effectiveness of these procedures is evaluated with five types of evidence: a) for the anchoring vignettes, the equivalence of the vignettes; b) the internal consistency of the factor or scale scores from the different procedures (i.e., Cronbach's alpha), which serves as a prerequisite psychometric check before conducting invariance tests; c) the effects of the different procedures on equivalence (as a baseline, we expect that each procedure would lead to some improvement of comparability over the raw continuous scores, given that all the procedures we test have been proposed to alleviate bias issues); d) the effects on the nomological networks of the self-assessed personality and value scores; and e) the effects on the external validity of country-level conscientiousness scores, which we operationalize by means of correlations with country-level development indices in line with previous research (Mõttus et al., 2012a, 2012b).

Method

Participants

Participants were 3,560 university students (27% males) with an age range of 16 to 37 years in 16 countries. The mean age of these participants was 22.19 years (SD = 2.29). Countries differed in affluence level and value dimensions such as collectivism and uncertainty avoidance, which are relevant for scale usage differences. The demographics are presented in Table 1. We initially aimed to collect 100 responses per country in order to suit CFA model testing of 10 items, and in most countries this was achieved except for Lebanon, Germany, Bulgaria, and Guatemala. The questionnaire and all analysis records except the dataset are available on OSF (identifier: DOI 10.17605/OSF.IO/UTKHY | ARK c7605/osf.io/utkhy).The dataset is hosted by the Royal Netherlands Academy of Arts and Sciences (https://doi.org/10.17026/dans-xbc-ecvr), and interested readers can obtain permission to access the files from the first author.

Procedures

Administration procedures were standardized with slight variations across countries, given local contextual differences. In countries with English as the native language or the main language of instruction in the university, the questionnaire was administered in English. In all other countries, the questionnaire was translated by two independent translators and convergence was sought to produce a final version. Collaborators in each country took responsibility in collecting data in the university where they work or are associated with. In Bulgaria, China, Indonesia, and Zambia, data were collected with paper and pencil in group settings, whereas in all other countries an online survey was administered. There is evidence that mode effects are very small in self-reports of SDR (He et al., 2015); therefore, we treated the different modes as interchangeable. The participation of all students was voluntary, although students in Canada, the Netherlands, South Africa, Singapore, Spain, and Turkey received course credits for their participation.

Measures

Anchoring vignettes. To keep the administration load and participant burden manageable, we chose to employ eight sets of anchoring vignettes, targeting conscientiousness (one vignette set for orderliness and one for industriousness), two value types (self-enhancement and openness to change), and the less explored ARS, ERS, MRS, and SDR. These vignettes were adapted from Mõttus et al. (2012a). Each set contained one high and one low trait-level vignette. These adapted vignettes were piloted together with the direct measures of response styles in China (n = 34) and in South Africa (n = 37). We estimated the internal consistency of ratings on high-level traits and on low-level traits separately (i.e., treating the eight high-level vignettes and the eight low-level vignettes as items of two separate scales). The Cronbach's alpha was above .50; suggestions from pilot respondents were taken into consideration to improve the vignettes. All vignettes were rated on a 5-point scale with different response options. For example, the vignettes for self-enhancement read:

Instruction: Read the description for each person below, and rate how much the statement "Being very successful is important to him. He hopes people will recognize his achievements" resembles him.

1 Already since childhood Bruno has wanted to achieve a lot in his life and he has worked a lot for it. Despite extreme poverty at home, he managed to get good education. Long hours at work have made him a very valued specialist and he has received ever better job offers.

2 Greg used to work as a salesman at a shop, but recently he asked to be a cleaner, because this would require less effort and shorter working hours. Although he lost a significant share of his salary, he is happy to have more free time to sit in front of the TV at home.

Self-report measures of ARS, ERS, MRS, and SDR. Self-report measures of ARS, ERS, and MRS, developed and validated in He and van de Vijver (2013), were further adapted based on the pilot study. In our previous study, self-report measures of response styles loaded on the same general response-style factor as the more frequently used indirect measures. Each style used balanced scales (i.e., half positively worded items and half negatively worded items) in an interrogative format (i.e., asking questions instead of rating on a statement) using semantic differentials. For example, each item had a different set of response options such as from *never* to *always*, *not important at all* to *extremely important*; this format has been shown to enhance cross-cultural comparability and to induce fewer response styles (e.g., Friborg, Martinussen, & Rosenvinge, 2006). Specifically, each scale comprised 10 questions and had five response anchors. A sample item for ARS reads: "Do you think it is good to always agree with others?"; a sample item for ERS reads: "Do you ever express an extreme idea, while your actual opinion is less extreme?; and a sample item for MRS reads: "How often do you give neutral opinions?"

Fifteen simplified and adapted items from the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960) were administered to tap into positive and negative impression management. All the SDR items were rated on a 5-point scale from 1 (*strongly disagree*) to 5 (*strongly agree*).

Personality. The Big Five personality scales (*Agreeableness*, *Conscientiousness*, *Extroversion*, *Openness*, and *Emotional Stability*) were measured with 50 items of the International Personality Item Pool (Goldberg et al., 2006) with response options ranging from 1 (*very inaccurate*) to 5 (*very accurate*).

Values. The four value dimensions (*Self-enhancement, Self-transcendence, Openness to Change*, and *Conservation*) were measured with the 21-item Portrait Values Questionnaire (Schwartz et al., 2001), with responses ranging from 1 (*does not resemble me at all*) to 5 (*very much resembles me*).

Background information, such as age, gender, and perceived social status, was collected at the beginning of the questionnaire. Responses to 45 other random questions of different response formats were also elicited at the end of the questionnaire, but not used in the current study. For the country-level validity analysis, we used GDP per capita, life expectancy from the World Bank database, and students' science achievement in the 2015 Programme for International Student Assessment.

Results

Analysis Strategy

The data analysis comprised four steps. In Step 1, where anchoring vignettes were the focus, the assumption of vignette equivalence was investigated in measurement invariance tests of ratings on vignettes; many studies of anchoring vignettes take this invariance for granted. In Step 2, the score transformation of target constructs including the five personality traits and the four value dimensions was the focus. The correction based on four different

statistical procedures (i.e., anchoring vignettes, GRS, parceling, and ipsatization) was performed, and the internal consistency for each set of scores was compared. Although internal consistency is not central in this study, the values of the internal consistencies in raw and corrected item scores may have implications for (or correspond to) the outcome of measurement invariance tests. In Step 3, measurement invariance of target scales of raw scores and corrected scores was compared using MGCFA in Mplus 7.1 (Muthen & Muthen, 1998-2012), where data were treated as continuous with robust ML estimation and (when possible) categorical with WLSMV estimation, respectively. In Step 4, the implications of procedures used in Step 3 were investigated by examining the correlations between personality traits and values. Each analysis step is described below in a separate section.

Missing value treatment. To account for missing values in the dataset, which would create difficulties in some of the procedures (e.g., rescaling based on anchoring vignettes can only be done with complete responses to the vignettes and to the target constructs, ipsatization involving the individual means of personality and values would result in missing in all items for respondents with a missing value on any personality or value items), a missing value analysis was first carried out. Less than 5% of missing were present in the data, yet they were not missing completely at random (Little's missing completely at random [MCAR] test: $\chi^2[70,949] = 84,665.99$, $p < .001$), so Expectation-Maximization imputation was used to replace all missing values (Schafer & Graham, 2002). The imputed values were then rounded to the nearest integers to facilitate later use of data as categorical.

Vignettes Equivalence

Quality of ratings on vignettes was first evaluated by computing the percentage of misorderings (i.e., inconsistent ordering) in vignette ratings in each set. Misorderings occur when a rating on a lower trait level is higher than a rating on a higher trait level. For instance, in rating vignettes describing persons running 5, 10, and 15 kilometers, respectively, on

degree of “being active in sports”, if respondents rate the hypothetical person running 5 kilometers as more active than the hypothetical person running 15 kilometers, this would be counted as a misordering. It suggests insufficient discrimination between the two vignettes, or respondents’ idiosyncratic interpretations of vignettes, which make these responses hard to analyze. For the eight sets of vignettes, the percentage of misorderings was 4% (orderliness), 30% (industriousness), 5% (self-enhancement) and 5% (openness to change), 6% (ARS), 11% (ERS), 13% (MRS), and 3% (SDR), respectively. In general, all vignette sets worked well, except the set for industriousness; the vignette about this aspect of conscientiousness was negatively worded as “to avoid duty”. The negation is likely to have caused confusion and required more complex cognition in the rating task. This set was removed from further analysis. Because the two value sets had a rather similar quality, only the self-enhancement set was retained for further analysis.

With the remaining six sets of ratings, the mean Cronbach’s alpha value was .653 (*SD* = .119) for the six high trait ratings (ranging from .324 in Bulgaria to .758 in Germany), and that of the six low trait ratings was .486 (*SD* = .286) ranging from -.155 in Germany to .693 in South Africa), indicating some consistency in ratings across content domains and a trait-like nature of the ratings, although domain-dependency is noticeable with these moderate internal consistency values and large variations existed in some countries. Next, to check the vignette equivalence assumption, ratings on the high trait vignettes from each of the six target constructs were subjected to MGCFA as a one-factor model; the same analysis was performed for the ratings on the six low trait vignettes, and for the deviance of ratings on high and low trait vignettes in the six target constructs, separately. If the assumed vignette equivalence holds, it would indicate that vignette evaluations are comparable across individuals and cultures. In operational terms, this means that ratings can differ across constructs, but there should be no measurement bias in responses to all high (or low) trait vignettes, and scalar

invariance should hold for the measurement of vignette ratings from different content domains (i.e., different target constructs). Given that the six ratings are in general either consistently high (for the high trait ratings) or low (for the low trait ratings), many categories had zero observations; so only the continuous model was used. Table 2 presents the results of the three MGCFAs. As shown in the model fit indexes, none of the ratings reached scalar invariance, whereas only the scale made of the six low trait level ratings reached metric invariance.[2] Clearly, vignettes equivalence was an untenable assumption in the current study of different response styles, personality, and values.

Procedures to Deal with Scale Usage Differences

Rescaling based on anchoring vignettes. Despite the observed vignette inequivalence, rescaling of target construct items based on dedicated vignettes was conducted. We briefly describe a nonparametric scoring approach here, also used in the present study. Such an approach rescales self-assessment responses (denoted as y) on the basis of responses of J ordered vignette questions (denoted as z_1 to z_j) to a single variable C (Equation 1) (King et al., 2004). With ratings on two vignettes ($J = 2$), the rescaled scores were also on a 5-point scale (i.e., $2J+1$), which were comparable to the raw responses. In cases of tied or inconsistently ordered vignette responses (e.g., $z_1 = z_2 = y$, or $z_2 < y = z_1$), the self-assessment responses can take a vector of possible values instead of one scalar value. For instance, if the comparisons of self-assessment y with two vignettes z_1 (lower trait level) and z_2 (higher trait level) show a pattern of $z_2 < y = z_1$, C may take any of the values from 2 to 5.

[2] We tried to establish partial equivalence of vignette ratings with MGCFAs by (1) excluding countries with low Cronbach's alpha in the vignette ratings, (2) only testing the assumption with the 6 countries using English as the survey, (3) treating cases with misorderings as missing in all countries, and (4) excluding vignettes that showed the largest loading variations. None of these analyses yielded scalar invariance for the high or low trait ratings, indicating the lack of equivalence is not easy to solve.

$$C = \begin{cases} 1 & if\ y < z1 \\ 2 & if\ y = z1 \\ 3 & if\ z1 < y < z2 \\ & \dots \\ 2J + 1 & if\ y > zj \end{cases}$$

The rescaling based on vignette ratings was carried out in the anchors package in R (Wand & King, 2007). As all sets of vignette ratings in use were positively formulated, the negatively worded items in target constructs were first reverse-coded to align with the rescaling. Items for each construct were rescaled based on the set of vignettes targeting the construct (personality or value). In cases of any violation (i.e., ties or misorderings), the rescaled responses had a range of possible values, and in the anchors package, the lowest (Cs) and the highest (Ce) possible rating could be produced. As the latter has been suggested to work better in enhancing comparability and validity (Kyllonen & Bertling, 2014), rescaled scores with the highest possible rating in case of violations were used as a proxy in this study. All rescaled responses had five-point responses.

GRS construction and correction. The scale scores of ARS, ERS, MRS, and SDR (combining the positive and negative impression management dimension) were used as indicators of the GRS in a Principal Component Analysis of the pooled sample data. A one-factor solution was supported (as the first four eigenvalues were 1.636, .969, .747, and .648, respectively). This factor explained 41% of the variance, with ERS (.74) and SDR (.39) loading positively, and ARS (-.62) and MRS (-.75) loading negatively. The factor score was extracted and used as GRS indicator in subsequent analyses. Next, each personality and values item was regressed on GRS, and the unstandardized residual score was taken as the GRS-corrected item response.

Ipsatization. Each respondent's mean of the 50 personality items (before reverse-coding) was calculated and subtracted from each personality item to produce ipsatized scores

of personality. Similarly, the individual mean of the 21 value items was calculated and subtracted from each value item to produce ipsatized scores of values.

Parceling. Parcels were calculated as the average item response of items forming the parcel.[3] As each of the five personality traits was measured by 10 items, four parcels were formed from these 10 items (in most cases, two parcels from positively worded items and two parcels from negatively worded items). For the four value dimensions, self-enhancement was measured by four items, so the parceling was the same as the original scale, whereas four parcels were calculated for each of the other three value dimensions. It should be acknowledged that these parcels are not unique, and replications with different combinations could be done. For instance, the combination of positively and negatively worded items into the same parcel may result in more equivalent parcels if bias would be related to the dimensionality or bias of the items. The current study focuses on only this set of parcels as a conservative test and an illustration of the impact of parceling.

Scale Consistencies

The internal consistency of each personality and value scale using raw scores and with the above mentioned procedures was computed. Table 3 presents the mean value and standard deviation of Cronbach's alpha across countries in these five sets of scores. We compared all correction procedures with the alpha values found for the raw scores using the Wilcoxon signed rank test. All pairwise differences were significant, meaning that the internal consistencies of the anchored responses were higher than those for the raw scores, whereas all other correction procedures revealed lower internal consistencies than those of the raw scores; *Z* scores were 2.67 for the difference between anchored and raw scores and -2.42, -2.68, and -2.54 for GRS, ipsatized and parcels, respectively (all *p*s < .05).

[3] The unidimensionality of each personality trait and value dimension was checked in a Principal Component Analysis with the pooled sample, and the scree plots supported the one-factor solution for each scale.

Measurement Invariance

With the four procedures introduced in the previous section, several sets of scores of personality and values were compared in measurement invariance testing with MGCFA: (1) raw responses, (2) anchored responses, (3) GRS-corrected scores, (4) ipsatized scores, and (5) item parcels. Nine unidimensional constructs were tested separately, including five personality traits (i.e., agreeableness, extraversion, conscientiousness, openness, and emotional stability) and four value dimensions (self-enhancement, self-transcendence, openness to change, and conservation). Therefore, the results speak to the equivalence of the individual factors, but not to the overall models of personality (i.e., Big Five) and values (i.e., Schwartz's two dimensions with four factors). The rationale is that equivalence of the overall models would be too high an aim or involve too complex models. For raw responses and anchored responses, the integer responses could be treated as both continuous and categorical, so both model methods were used. For the ipsatized and parcel scores, it was not possible to use the categorical approach due to the continuous nature of these scores. For the categorical approach, given the missing cells in some categories in some countries, the 5-point responses were collapsed to three categories (the original 1 and 2 recoded to 1, 3 recoded to 2, and 4 and 5 recoded to 3). Table 4 presents a summary of CFI values in each construct and for each procedure. The full model fit statistics (including Chi-square statistics, degree of freedom, CFI, TLI, and RMSEA) are provided in the supplementary materials. It should be noted that the model fit criteria in treating the data as continuous and categorical are not directly comparable given the different modeling approaches. Thus, the levels of comparability established are independently evaluated from either the categorical or continuous approach.

In the continuous models, constructs measured with raw responses in general showed very poor model fit even at configural level, except for Self-Enhancement which showed adequate configural invariance across all 16 countries. Compared with raw responses,

ipsatization led to lower levels of comparability in all constructs, as indicated by lower CFI values, whereas fit indexes in GRS-corrected scores remained largely intact, suggesting that correction for GRS did not improve or jeopardize the comparability of scales. It is noteworthy that anchored scores and parceling in most cases enhanced comparability, especially at configural and metric levels. Specifically, with anchored scores, the four value dimensions reached metric invariance, and agreeableness and extraversion reached marginal metric invariance. With parcels, extraversion, openness, self-transcendence and conservation showed metric invariance, and agreeableness, emotional stability and self-enhancement showed configural invariance. However, in no case was scalar invariance reached.

In the categorical models, the collapse of categories from five to three did not solve all model identification problems; a further score collapse would have been needed for some scales, which was not pursued because of the huge loss of information implied in these score reductions. With raw responses, all identified models reached configural invariance except openness and agreeableness (not estimated). In anchored responses, all identified models reached configural invariance, and agreeableness reached metric invariance.

In summary, the different procedures proposed to account for scale usage differences and to enhance comparability in self-report personality and value data had different effects; ipsatization and GRS-correction had small effects on comparability, whereas the use of anchoring vignettes and parceling had larger effects and increased comparability; yet, none of these approaches resulted in scalar invariance for any scale. Treating data as categorical tended to yield the highest level of invariance, but categorical models cannot be applied when certain categories have zero observations. It seems that anchoring vignettes and parceling in the continuous models had the most merit, so these two were combined (i.e., creating parcels with anchored responses and testing measurement invariance of these anchored parcels in

continuous models). The last column in Table 4 (i.e., 6 Anchored-Parcel-Continuous) presents the CFI values in such a case. All scales reached metric invariance.

Nomological Networks

The implications of the different procedures were investigated in correlational analyses of personality traits and values with different procedures across cultures. The central question here is whether the nomological network of a trait or value changes as a function of score corrections. This approach was adopted because without scalar invariance, none of the scale scores could be directly compared across cultures; yet with some scales reaching metric invariance, associations among scales can be validly compared across cultures.

Firstly, factor scores of each scale were produced in a CFA model with respondents from all countries (using each of the procedures in Step 3). Correlations of the eight sets of factor scores of each construct were computed to check if different procedures produce different score patterns. Table 5 presents the intercorrelations of conscientiousness across the eight sets of scores. When the same data were modeled as continuous or categorical (e.g., raw scores, anchored scores), the resulting factor scores had extremely strong correlations (in all cases, correlations above .99), indicating that treating data as continuous and categorical produced similar factor scores. Factor scores from raw scores (procedures 1.1, 1.2), GRS corrected scores (procedure 3), ipsatized scores (procedure 4), and parcel scores (procedure 5) showed strong correlations (in .80s – .90s range), whereas anchored scores and anchored parcel scores (procedures 2.1, 2.2, and 6) showed weaker correlations with the first five sets of factor scores (.40s – .50s range). The intercorrelations of all other constructs followed a similar patterning, as shown in the correlations of the first set of factor scores with all other procedures in Table 6.

The correlations between personality traits and values we studied were agreeableness with self-transcendence, extraversion with self-enhancement, conscientiousness with

conservation, and openness with openness to change, respectively. These personality traits and values have been found to be consistently, positively correlated (Parks-Leduc, Feldman, & Bardi, 2015). Table 7 presents the median correlations across countries with each procedure. The raw, GRS-corrected, and parceled scores seemed to produce rather similar correlations, whereas anchoring and ipsatization resulted in overall weaker correlations. We again compared the correlations using the Wilcoxon signed rank test. None of the pairwise differences between the uncorrected (raw-continuous) scores and the other seven sets of scores was statistically significant; *Z* scores were -1.83 for the difference between raw-categorical and raw-continuous scores and -1.83, -1.83, -.37, -1.46, -.73 and -1.83 for anchor-continuous, anchor-categorical, GRS, ipsatized, parcels and anchored parcels, respectively (all *p*s >.05). To summarize, none of the procedures significantly changed the correlations.

Country-Level Validity

We linked aggregated raw and corrected scores of conscientiousness at country level with GDP per capita, life expectancy, and students' science achievement in the 2015 Programme for International Student Assessment, using Spearman's rank-order correlation. It could be expected that an effective correction procedure would result in more positive (or at least fewer negative) correlations of country-level conscientiousness with these external validity measures indicative of human development (Mõttus et al., 2012b). Table 8 presents the correlations based on 1000 bootstrapped samples (the bootstrapping was applied given the small number of observations at country level). Procedures involving anchoring vignettes (2.1, 2.2, and 6) and ipsatization (4) did not improve the validity, but instead they seemed to decrease country-level validity. Parceling did not make much difference, whereas GRS correction seemed to produce country-level correlations that were more in line with expectations.

Discussion

Data incomparability in self-report Likert scales presents a perennial challenge to drawing valid comparative inferences and utilizing cross-cultural data for interventions or policy making. This study aimed to shed light on whether different design and data analysis procedures that can presumably correct for scale usage differences indeed enhance data comparability, and the implications of these procedures. Using data collected among university students in 16 countries, we investigated the effects of anchoring vignettes, the general response style, data ipsatization, and item parceling on personality and value measures. The main conclusions were as follows: (1) ratings on vignettes from different cultures were not bias-free, and so vignette equivalence as a working assumption may not hold; (2) as for the internal consistency of scales, internal consistency was enhanced with anchored scores, and reduced somewhat by GRS correction, parceling and, especially, ipsatization; (3) as for measurement invariance across cultures, anchoring vignettes and item parceling seemed to be able to improve the level of comparability to some extent, response style correction left the results from raw responses intact, and ipsatization resulted in more incomparability of value and personality scales, although in no case was scalar invariance achieved; (4) when comparing treating Likert-scale data as continuous and categorical, the latter in general showed higher levels of comparability although identifiability of such models can be challenged by the occurrence of empty cells; (5) a comparison of the factor scores of scales extracted from different procedures showed much similarity, and the associations among factor scores of different constructs were rather similar across procedures; and (6) correcting for the GRS was the only procedure that suggested an increase in the external validity of country-level scores on conscientiousness. We discuss each finding separately and provide recommendations for cross-cultural comparative studies on personality and beyond.

Drawbacks of Anchoring Vignettes

Without vignette equivalence, the golden standard implied by the use of anchors is on shaky ground (Hopkins & King, 2010). This assumption is often taken for granted, and few studies have explicitly tested for its tenability (Ferrer-i-Carbonell, van Praag, & Theodossiou, 2011). Using MGCFA, we tested the often tacitly assumed equivalence of vignette ratings and of the correction factor (operationalized as the deviance of the high and low trait level rating) in six sets of vignette rating data targeting different constructs; we found that neither the vignettes nor the correction factor derived from the vignettes were equivalent across countries. Although our study involved a relatively small number of vignettes and our findings await replication, there is cause for concern. Contrary to common practice in working with anchoring vignettes, it is counterintuitive to simply assume that vignettes would be immune to comparability challenges when there is so much evidence of poor comparability of all kinds of response scale formats in large-scale surveys. Our finding has two immediate implications: Firstly, the rescaled scores based on vignette ratings are not free of bias, although the rescaled scores may be less biased compared with the raw scores. Secondly, to the extent that a correction factor obtained from a certain set of vignettes is not invariant, it cannot be applied to different constructs either; so, vignettes targeting one construct should not be used to rescale raw scores of other constructs, and different sets of vignettes are not interchangeable.

Vignette inequivalence could be due to the fact that descriptions of hypothetical persons are situated in specific contexts, and these contextual cues are interpreted differently across cultures. This is a dilemma in vignette designing: short and concise statements could contain fewer triggers of inequivalence, yet vignettes require elaboration (e.g., providing contexts or examples of manifestation of trait level). Another reason for vignette inequivalence may be the misorderings in the datasets. Misorderings are logically uninterpretable, and compromise data quality. This was confirmed in an additional MGCFA when misorderings were recoded as missing values. In this case, the deviance scores (but not

the low- and high-trait scores) showed an acceptable metric invariance, suggesting the stability of the structure and factor loadings of the vignettes correction factor across domains, yet some uniform item bias continued to be present. This finding again speaks to the recommendation not to use one set of vignettes to rescale various domains not targeted in the vignettes (Primi, Zanon, Santos, De Fruyt, & John, 2016).

Internal Consistency

With raw responses, internal consistencies of all five personality traits and four value dimensions were acceptable. There were changes associated with various correction procedures. Specifically, rescaled scores based on anchoring vignettes had a higher internal consistency, presumably due to the fact that all items in one scale were rescaled according to one set of vignette ratings, which boosts the inter-item correlations by employing one correction factor. GRS-correction slightly lowered the internal consistency, indicating that GRS tends to slightly lower the item intercorrelations. Ipsatized scores showed poorer internal consistency, and the impact was stronger for value dimensions than for personality traits. Some reduction in Cronbach's alphas could in principle be expected given that all the constructs are interrelated. As there was a smaller number of items in parcels compared with that of raw responses, the internal consistency values was somewhat lower; yet all scales still had a Cronbach's alpha value above .60, which proves that parceling maintains levels of internal consistency. All in all, although high internal consistency is often considered desirable, it may not be a good criterion for internal consistency or validity (compared with test-retest internal consistency), because high internal consistency may result from scale items permeated by common method effects, which may decrease comparability of scales across cultures (McCrae, 2015).

Effects of Correction Procedures on Measurement Invariance

We compared MGCFA outcomes with different procedures for each value and personality construct. None of the procedures solved the problem of a lack of scalar invariance for personality and values across the 16 countries. Anchoring vignettes and parceling improved comparability, GRS-correction resulted in very similar fit indexes as the raw scores, and ipsatization resulted in deteriorated comparability. The incremental value of anchoring vignettes lies in the reduction of reference group effects in raw responses, although this procedure is not a cure-all, given the problematic nature of the assumption of vignette equivalence. The better model fit with parcels could be due to item bias canceling out when forming the parcels, but also the fewer parameters in the models to be estimated. The rather limited impact of GRS correction could be explained by the nature of GRS as a self-presentation style embedded in personality and values that impacts all self-reports (Billiet & Davidov, 2008; He & van de Vijver, 2015b). As a consequence, correcting for GRS does not change the structure or metrics of raw responses.

In comparing two modeling methods in MGCFA (continuous versus categorical), categorical models have the advantage of requiring less restrictive assumptions about the data distribution, and with more item parameters to estimate, the models have more flexibility to find a better model fit (Rutkowski & Svetina, 2016). At the same time, the increased number of parameters means lower power to detect lack of invariance, so the improved fit may to some extent be artefactual. We confirmed better model fit and higher levels of comparability in categorical models. The drawback of this categorical approach is that it requires that every group has sufficient observations in every category.

Nomological Networks

In comparing factor scores across different procedures, continuous and categorical models of the same data produced very similar scores, even though measurement invariance tests tend to favor categorical models. Parceling scores are most strongly correlated with raw

scores followed by GRS-corrected scores, ipsatized scores and anchored scores, indicating that anchor vignettes make most drastic score corrections. The correlations between values and personality per country seem rather stable across procedures, although anchored scores tend to show slightly lower correlations (although not significantly different from raw scores), possibly because the rescaling based on different target constructs removed some common method variance in personality and value data. In some cases, ipsatization produced negative correlations (nonsignificant) between values and personality. This suggests that ipsatization could remove valid substance in values and personality, and perhaps results in over-correction. This is in line with the measurement invariance test in Step 3, suggesting that ipsatization jeopardizes the measurement comparability and perhaps distorts scale correlations further.

Country-Level Validity

Our study did not fully replicate Mõttus et al.'s (2012a, 2012b) results in testifying the validity enhancement with anchoring vignettes, which could be due to the specific student samples in too few countries. However, our findings were in line with the Mõttus et al. (2012b) study, to the extent that correcting for response styles, although done in a different way, results in more intuitive correlations.

Practical Recommendations

Taking into account all the different analyses employed, we conclude that none of the procedures tested in this study can fully resolve the issue of incomparability of data in large-scale cross-cultural studies on personality and values. We acknowledge that with more cultures involved, full measurement invariance is difficult to reach, as there is more cultural variation at the item level than at the structural level (i.e., configural invariance known as the psychic unity of humankind is easier to reach). However, if the cultural variations are minor and can to some extent be corrected for, then valid comparisons can still be made.

For the design-based approaches, *anchoring vignettes* have attractive features, because the rescaling has the potential to enhance comparability somewhat. However, there is a need to test the tenability of the two assumptions, notably vignette equivalence, before using them in practice (Ferrer-i-Carbonell et al., 2011).

In terms of analysis, *parcels,* when used as latent trait indicators in the CFA framework, could help to reach higher levels of comparability. (We do not further consider here the conceptual arguments against parceling; see, e.g., Marsh et al., 2013.) Moreover, factor scores extracted from parcels do not deviate from raw scores in unidimensional scales, suggesting that parcels help simplify the measurement model without distorting the meaning of the construct: With item bias canceling each other out within parcels, parcels tend to be more comparable indicators of a scale.

For modeling methods, *categorical MGCFA* seems to be a better method compared with continuous models, if model identification is not an issue (i.e., given non-zero observations in each category in each group). To remedy the identification problem in categorical models, several options can be tried out: (1) collapsing adjacent categories (if many categories are available, and only few categories have missing values spreading in different groups), (2) removing groups or items with missing categories (if there are only a few groups with missing categories spreading in different items, or if only a few items out of many have missing categories across groups), and (3) creating (artificial) cases with missing categories in groups in the dataset for identification (in large sample sized study where the artificial cases do not impact on the general structure of data).

Correcting for GRS does not make much difference in either internal consistency or measurement comparability; therefore, if the research focus is not response styles themselves, collecting external response style data does not seem to be helpful.

Ipsatization, which is perhaps the most popular of the studied techniques in personality, is not as successful in our data as in previous studies. Rammstedt et al. (2013) focused on the agreement of data with an ideal structure, and found the most benefit of ipsatization in individualistic cultures. In contrast, we found that ipsatization might amount to over-correction by lowering the internal consistency and measurement invariance. Therefore, the last recommendation is to be very cautious in using *ipsatization* for cross-cultural comparisons.

Finally, it is difficult to conclude that higher comparability (and/or higher internal consistency) indicates higher validity. Global recommendations as to which procedure is preferred depend a lot on the research goals and the specific criterion used to judge these procedures. For individual-level measurement-invariance testing, it seems that anchoring vignettes and parceling would produce better outcomes, whereas for country-level correlational analyses, GRS correction seems to be more productive. We do not have sufficient evidence to favor any particular procedure for all purposes.

Limitations and Further Directions

This study has a few limitations. Firstly, we only have student samples from 16 countries, and in most countries there were more female than male students. Furthermore, with Bulgaria, Germany, Guatemala, and Lebanon each having a sample size smaller than 100, there was concern for power in the MGCFAs. Two sets of analyses were carried out to ensure the findings were not affected by the small sample sizes in these countries. First, we replicated the MGCFAs without these four countries, and did not find a different patterning. Second, we carried out the MGCFAs with pseudo-countries of the same sample size per country by randomly selecting cases from the whole sample pool. We found much better model fit compared with the analyses of real countries, and in most cases the scalar invariance model fit

almost as well as the configural model. These results suggest that the smaller sample sizes in four countries have not substantially affected our main conclusions.

Another limitation is that vignettes in this study involved studying, job performance, relationship, and personal history; it is possible that some students could not relate to certain situations personally, whereas a general population may rate these vignettes as more personally relevant. Moreover, the anchored scores of personality and values were based on a single factor from both domains, which may result in some confounding. More sets of vignettes targeting each trait and value may be of help. Finally, future studies could test novel ways for dealing with response style not addressed here, such as neutral formulation of items (Bäckström, Björklund, & Larsson, 2014).

Conclusion

We assessed the effects of anchoring vignettes, response-style correction, data ipsatization, and item parceling on the equivalence of personality and value measures in data from 16 countries. Although none of the procedures was a clear winner across our criteria, we highlighted the effects of the different procedures on scale usage differences in cross-cultural research. We encourage researchers to rigorously test data comparability, employ innovative design features (e.g., Kyllonen & Bertling, 2014; Revilla, Saris, & Krosnick, 2014), and utilize appropriate psychometric methods (e.g., Rutkowski & Svetina, 2016; van de Schoot et al., 2013), including different procedures for assessing response styles to enhance the quality of data and inferences. Our study does not suggest that there is a magic bullet to achieve scalar invariance. Yet, a combination of design and analysis procedures, linked to the aims of the study, can go a long way.

References

Bäckström, M., & Björklund, F. (2014). Social desirability in personality inventories: The nature of the evaluative factor. *Journal of Individual Differences, 35*, 144-157. doi:10.1027/1614-0001/a000138

Bäckström, M., Björklund, F., & Larsson, M. R. (2014). Criterion validity is maintained when items are evaluatively neutralized: Evidence from a full-scale Five-factor Model inventory. *European Journal of Personality, 28*, 620–633. doi:10.1002/per.1960

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*, 78-102. doi:10.1207/S15328007SEM0901_5

Barrett, P. T., Petrides, K., Eysenck, S. B. G., & Eysenck, H. J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences, 25*, 805–819. doi:10.1016/S0191-8869(98)00026-9

Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research, 36*, 542-562. doi:10.1177/0049124107313901

Bilsky, W., Janik, M., & Schwartz, S. H. (2011). The structural organization of human values—Evidence from three rounds of the European Social Survey (ESS). *Journal of Cross-Cultural Psychology, 42*, 759–776. doi:10.1177/0022022110362757

Bou Malham, P., & Saucier, G. (2016). The conceptual link between social desirability and cultural normativity. *International Journal of Psychology, 51*, 474-480. doi:10.1002/ijop.12261

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. doi:10.1207/s15328007sem0902_5

Church, A. T., Anderson-Harumi, C. A., del Prado, A. M., Curtis, G. J., Tanaka-Matsumi, J., Valdez Medina, J. L., ... White, F. A. (2008). Culture, cross-role consistency, and adjustment: Testing trait and cultural psychology perspectives. *Journal of Personality and Social Psychology, 95*, 739-755. doi:10.1037/0022-3514.95.3.739

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly, 72*, 420-445. doi:10.1093/poq/nfn035

De Raad, B., Barelds, D. P. H., Timmerman, M. E., De Roover, K., Mlačić, B., & Church, A. T. (2014). Towards a pan-cultural personality structure: Input from 11 psycholexical studies. *European Journal of Personality, 28*, 497–510. doi:10.1002/per.1953

Eigenhuis, A., Kamphuis, J. H., & Noordhof, A. (2015). Personality differences between the United States and the Netherlands: The influence of violations of measurement invariance. *Journal of Cross-Cultural Psychology, 46*, 549–564. doi:10.1177/0022022115570671

Ferrer-i-Carbonell, A., van Praag, B. M. S., & Theodossiou, I. (2011). *Vignette equivalence and response consistency: The case of job satisfaction*. IZA Discussion Paper (No. 6174).

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology, 35,* 263-282. doi:10.1177/0022022104264122

Fischer, R. (2014). What values can (and cannot) tell us about individuals, society and culture. In M. J. Gelfand, C. Chiu, & Y. Hong (Eds.), *Advances in culture and psychology* (Vol. 4, pp. 218–272). Oxford, United Kingdom: Oxford University Press.

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences, 40,* 873-884. doi:10.1016/j.paid.2005.08.015

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84-96. doi:10.1016/j.jrp.2005.08.007

He, J., & van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences, 55,* 794-800. doi:10.1016/j.paid.2013.06.017

He, J., & van de Vijver, F. J. R. (2015a). Effects of a general response style on cross-cultural comparisons: Evidence from the Teaching and Learning International Survey. *Public Opinion Quarterly, 79,* 267-290. doi:10.1093/poq/nfv006

He, J., & van de Vijver, F. J. R. (2015b). Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences, 31,* 129-134. doi:10.1016/j.paid.2014.09.009

He, J., van de Vijver, F. J. R., Domínguez Espinosa, A., Abubakar, A., Dimitrova, R., Adams, B., . . . Villieux, A. (2015). Socially desirable responding: Enhancement and denial in 20 countries. *Cross-Cultural Research, 49,* 227–249.

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly, 74,* 201-222. doi:10.1093/poq/nfq011

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*, 332–346. doi:10.1177/1088868310361240

Jürges, H., & Winter, J. (2013). Are anchoring vignettes ratings sensitive to vignette age and sex? . *Health Economics, 22,* 1-13. doi:10.1002/hec.1806

Kapteyn, A., Smith, J. P., Van Soest, A., & Vonkova, H. (2011) Anchoring vignettes and response consistency. *Vol. RAND working paper WR-840*: RAND.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*, 191-207. doi:10.1017/S000305540400108X

Kyllonen, P. C., & Bertling, J. J. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277-286). Boca Raton, FL: CRC Press.

Leung, K., Au, A., Huang, X., Kurman, J., Niit, T., & Niit, K.-K. (2007). Social axioms and values: A cross-cultural examination. *European Journal of Personality, 21*, 91–111. doi:10.1002/per.615

Leung, K., Lam. B. C. P., Bond, M. H., Conway, L. G., III, Gornick, L. J., Amponsah, B., ... Zhou, F. (2012). Developing and evaluating the Social Axioms Survey in eleven countries: Its relationship with the Five-Factor Model of personality. *Journal of Cross-Cultural Psychology, 43,* 833–857. doi:10.1177/0022022111416361

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right: Camouflaging misspecification with item parcels in CFA models. *Psychological Methods, 18,* 257-284. doi:10.1037/a0032773

McCrae, R. R. (2015). A more nuanced view of internal consistency. *Personality and Social Psychology Review, 19,* 97-112. doi:10.1177/1088868314541857

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*, 882-888. doi:10.1037/0022-006X.51.6.882

McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology, 88*, 547–561. doi:10.1037/0022-3514.88.3.547

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*, 450–470. doi:10.1037/a0019216

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515. doi:10.1207/S15327906MBR3903_4

Mõttus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., ... Tseung, C. N. (2012a). Comparability of self-reported conscientiousness across 21 countries. *European Journal of Personality, 26,* 303–317. doi:10.1002/per.840

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., . . . Johnson, W. (2012b). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin, 38,* 1423-1436. doi:10.1177/0146167212451275

Muthen, L. K., & Muthen, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthen & Muthen.

Nye, C. D., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality, 42,* 1524-1536. doi:10.1016/j.jrp.2008.07.004

OECD. (2014). *TALIS 2013 technical report*. Paris, France: OECD publishing.

Parks-Leduc, L., Feldman, G., & Bardi, A. (2015). Personality traits and personal values: A meta-analysis. *Personality and Social Psychology Review, 19,* 3-29. doi:10.1177/1088868314538548

Paulhus, D. L. (1991). Measurement and control of response biases. In J. Robinson, P. Shaver & L. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17-59). San Diego, CA: Academic Press.

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes: Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *European Journal of Psychological Assessment, 32,* 39-51. doi:10.1027/1015-5759/a000336

Rammstedt, B., Kemper, C. J., & Borg, I. (2013). Correcting big five personality measurements for acquiescence: An 18-country cross-cultural study. *European Journal of Personality, 27*, 71-81. doi:10.1002/per.1894

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research, 43*, 73-97. doi:10.1177/0049124113509605

Rohling, M. L., Larrabee, G. J., Greiffenstein, M. F., Ben-Porath, Y. S., Lees-Haley, P., & Green, P. (2010). A misleading review of response bias: Comment on McGrath, Mitchell, Kim, and Hough (2010). *Psychological Bulletin, 137*, 708–712. doi:10.1037/a0023327

Rossier, J., Aluja, A., Blanch, A., Barry, O., Hansenne, M., Carvalho, A. F., ... Karagonlar, G. (2016). Cross-cultural generalizability of the Alternative Five-factor Model using the Zuckerman–Kuhlman–Aluja Personality Questionnaire. *European Journal of Personality, 30*, 139–157. doi:10.1002/per.2045

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*, 31-57. doi:10.1177/0013164413498257

Rutkowski, L., & Svetina, D. (2016). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education*. doi:10.1080/08957347.2016.1243540

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63*, 506–516. doi:10.1207/s15327752jpa6303_8

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177. doi:10.1037/1082-989X.7.2.147

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of Big Five personality traits. *Journal of Cross-Cultural Psychology, 38,* 173-212. doi:10.1177/0022022106297299

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology, 25,* 1–65. doi:10.1016/S0065-2601(08)60281-6

Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality, 38,* 230–255. doi:10.1016/S0092-6566(03)00069-2

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology, 32,* 519-542. doi:10.1177/0022022101032005001

Schwartz, S. H., & Sagiv, L. (1995). Identifying culture-specifics in the content and structure of values. *Journal of Cross-Cultural Psychology 26,* 92-116. doi:10.1177/0022022195261007

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology, 35,* 50-61. doi:10.1177/0022022103260380

Smith, P. B. (2011). Communication styles as dimensions of national culture. *Journal of Cross-Cultural Psychology, 42,* 216-233. doi:10.1177/0022022110396866

Thalmayer, A. G., & Saucier, G. (2014). The Questionnaire Big Six in 26 nations: Developing cross-culturally applicable Big Six, Big Five and Big Two inventories. *European Journal of Personality, 28,* 482–496. doi:10.1002/per.1969

van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and

self-confidence across countries *Journal of Cross-Cultural Psychology, 43*, 1205-1228. doi:10.1177/0022022111428083

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 770. doi:10.3389/fpsyg.2013.00770

van de Vijver, F. J. R., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality, 69*, 1007-1031. doi:10.1111/1467-6494.696173

Wand, J., & King, G. (2007). Anchoring vignetttes in R: A (different kind of) vignette. Retrieved from http://wand.stanford.edu/anchors/doc/anchors.pdf

Table 1

Demographics of Participants

Country	Sample Size	Mean Age (*SD*)	% of Males	Language
Bulgaria	80	20.34 (1.20)	16.25	Bulgarian
Canada	431	21.77 (2.54)	24.88	English
China	309	20.76 (1.01)	12.30	Chinese
Germany	95	25.01 (4.74)	20.00	German
Guatemala	94	23.82 (4.02)	32.98	Spanish
Indonesia	403	22.32 (1.54)	30.02	English
Lebanon	74	22.89 (2.05)	25.68	Arabic
Lithuania	259	23.07 (2.76)	13.13	Lithuanian
Mexico	163	21.68 (2.23)	28.83	Spanish
Netherlands	206	21.63 (1.84)	20.87	Dutch
Romania	215	22.46 (2.39)	27.10	Romanian
Singapore	275	23.03 (1.30)	33.58	English
South Africa	306	21.62 (2.03)	32.89	English
Spain	127	21.83 (1.44)	17.46	Spanish
Turkey	223	22.42 (2.46)	39.64	Turkish
Zambia	300	22.20 (2.38)	40.20	English

Table 2

Measurement Invariance Testing of Vignette Ratings

Low Trait Ratings	χ^2	*df*	CFI	RMSEA
Configural	232.898	144	.949	.053
Metric	443.589	219	.948	.064
Scalar	1031.74	294	.581	.116
High Trait Ratings				
Configural	314.064	144	.920	.073
Metric	451.899	219	.890	.069
Scalar	1480.60	294	.441	.135
Deviance High and Low Trait Rating				
Configural	268.485	144	.951	.062
Metric	478.675	219	.899	.073
Scalar	1384.65	294	.575	.129

Table 3

Mean (and Standard Deviation) of Cronbach's Alpha Values for Each Scale Across Countries

Scale	1 Raw	2 Anchored	3 GRS-Corrected	4 Ipsatized	5[a] Parcels
Agreeableness	.72 (.08)	.87 (.03)	.71 (.08)	.57 (.13)	.66 (.10)
Extraversion	.82 (.09)	.89 (.05)	.80 (.08)	.71 (.09)	.78 (.10)
Conscientiousness	.77 (.05)	.86 (.03)	.76 (.05)	.64 (.15)	.74 (.06)
Openness	.73 (.06)	.87 (.05)	.71 (.05)	.59 (.17)	.69 (.10)
Emotional Stability	.80 (.09)	.89 (.05)	.80 (.09)	.73 (.12)	.74 (.12)
Self-Enhancement	.69 (.05)	.80 (.03)	.69 (.05)	.49 (.14)	.69 (.05)
Self-Transcendence	.68 (.09)	.82 (.03)	.67 (.09)	.35 (.21)	.65 (.10)
Openness to Change	.71 (.05)	.83 (.03)	.69 (.05)	.36 (.19)	.66 (.08)
Conservation	.64 (.10)	.79 (.07)	.63 (.10)	.33 (.15)	.60 (.11)

[a]Number of parcels equals 4 in all scales.

Table 4

Comparative Fit Index (CFI) Values of Measurement Invariance Testing of Target Constructs

Construct	Models	1.1 Raw Continuous	1.2 Raw Categorical	2.1 Anchored Continuous	2.2 Anchored-Categorical	3 GRS-Corrected Continuous	4 Ipsatized Continuous	5 Parcel Continuous	6 Anchored Parcel Continuous
Agreeableness	Configural	nc	ec	.903	.940	.804	nc	.907	.956
	Metric	.709		.882	.946	.709	.586	.842	.931
	Scalar	.389		.764	.908	.388	.050	.424	.784
Extraversion	Configural	.854	.937	.897	.949	.841	.728	.958	.972
	Metric	.826	.901	.888	.938	.811	.700	.949	.962
	Scalar	.645	.849	.803	.922	.613	.376	.758	.870
Conscientiousness	Configural	.792	.895	.852	.925	.791	nc	nc	.909
	Metric	.691	.829	.826	.912	.683	.572	.799	.893
	Scalar	.492	.767	.739	.892	.482	.294	.628	.813
Openness	Configural	.725	.843	.857	.926	.702	nc	.981	.989
	Metric	.684	.845	.852	.930	.659	.544	.961	.975
	Scalar	.531	.800	.792	.919	.492	.321	.791	.917
Emotional Stability	Configural	.752	.935	.853	.942	.750	.669	.965	.973
	Metric	.697	.911	.831	.932	.697	.617	.902	.951
	Scalar	.558	.875	.757	.923	.560	.427	.734	.856
Self-enhancement	Configural	.943	.973	.957	.981	.941	nc	.943	.957
	Metric	.905	nc	.947	.959	.899	.768	.905	.947
	Scalar	.621	.817	.817	.928	.608	.084	.621	.817
Self-transcendence	Configural	nc	ec	.982	ec	nc	nc	.965	.987
	Metric	.944		.976		.945	.766	.954	.978
	Scalar	.716		.879		.726	0	.727	.877
Openness to Change	Configural	.850	.932	.944	.969	.843	nc	.892	.959
	Metric	.817	nc	.927	.958	.804	.573	.860	.938
	Scalar	.554	.823	.824	.930	.542	0	.564	.822

Conservation	Configural	.888	.945	.941	.965	.887	nc	.930	.975
	Metric	.867	nc	.930	.928	.866	.681	.913	.964
	Scalar	.385	nc	.766	.884	.386	0	.562	.846

Note. nc stands for non-convergence in the model. ec in categorical models stands for empty cells that make estimation impossible.

Table 5

Intercorrelations of Factor Scores of Conscientiousness in Different Procedures across Countries

	1.1	1.2	2.1	2.2	3	4	5	6
1.1 Raw-continuous	1							
1.2 Raw-categorical	.995**	1						
2.1 Anchored-continuous	.531**	.529**	1					
2.2 Anchored-categorical	.535**	.534**	.997**	1				
3 GRS-corrected	.984**	.978**	.527**	.530**	1			
4 Ipsatized	.808**	.806**	.463**	.473**	.832**	1		
5 Parcels	.987**	.984**	.524**	.531**	.970**	.808**	1	
6 Anchored Parcels	.538**	.537**	.997**	.996**	.533**	.474**	.539**	1

**$p < .01$.

Table 6

Correlations of Factor Scores from Raw Scores (Continuous Model) with Factor Scores from All Other Procedures

	Agreeableness	Extraversion	Openness	Emotional Stability	Self-Enhancement	Self-Transcendence	Openness to Change	Conservation
1.2 Raw Categorical	.988**	.997**	.995**	.994**	.996**	.992**	.995**	.997**
2.1 Anchored Continuous	.402**	.587**	.477**	.619**	.622**	.537**	.573**	.581**
2.2 Anchored Categorical	.396**	.594**	.484**	.623**	.622**	.529**	.567**	.578**
3 GRS-corrected	.992**	.945**	.941**	.995**	.986**	.989**	.959**	.995**
4 Ipsatized	.730**	.863**	.766**	.879**	.790**	.681**	.676**	.660**
5 Parcels	.988**	.991**	.973**	.985**	1.00**	1.00**	.980**	.985**
6 Anchored Parcels	.401**	.578**	.474**	.613**	.622**	.538**	.561**	.574**

**$p < .01$.

Table 7

Median Correlations between Values and Personality Traits across Countries

Procedure	Agreeableness and Self-Transcendence	Conscientiousness and Conservation	Extraversion and Self-Enhancement	Openness and Openness to Change
1.1 Raw-continuous	.500	.261	.197	.387
1.2 Raw-categorical	.521	.275	.198	.391
2.1 Anchored-continuous	.242	.154	.105	.149
2.2 Anchored-categorical	.226	.148	.111	.154
3 GRS-corrected	.502	.283	.176	.326
4 Ipsatized	.294	-.063	.204	.098
5 Parcels	.489	.244	.220	.348
6 Anchored Parcels	.242	.149	.109	.141

Table 8 *Country-Level Spearman's Rank Order Correlations between Conscientiousness and External Validity Measures across Countries*

Spearman's rho	GDP per capita (N = 16)	Life Expectancy (N = 16)	PISA 2015 Science Achievement (N = 13)
1.1 Raw-continuous	-.256	-.376	.099
1.2 Raw-categorical	-.259	-.394	.088
2.1 Anchored-continuous	-.503*	-.353	-.060
2.2 Anchored-categorical	-.488	-.350	-.104
3 GRS-corrected	-.118	-.247	.297
4 Ipsatized	-.547*	-.491	-.099
5 Parcels	-.306	-.362	.104
6 Anchored Parcels	-.506*	-.332	-.033

Note. * p < .05. The bootstrap results are based on 1000 bootstrap samples.