The Effect of Response Style on Self-Reported Conscientiousness Across 20 Countries

René Mõttus, Jüri Allik, Anu Realo, Jérôme Rossier, Gregory Zecca, Jennifer Ah-Kion, Dénis Amoussou-Yéyé, Martin Bäckström, Rasa Barkauskiene, Oumar Barry, Uma Bhowon, Fredrik Björklund, Aleksandra Bochaver, Konstantin Bochaver, Gideon de Bruin, Helena F. Cabrera, Sylvia Xiaohua Chen, A. Timothy Church, Daouda Dougoumalé Cissé, Donatien Dahourou, Xiaohang Feng, Yanjun Guan, Hyi-Sung Hwang, Fazilah Idris, Marcia S. Katigbak, Peter Kuppens, Anna Kwiatkowska, Alfredas Laurinavicius, Khairul Anwar Mastor, David Matsumoto, Rainer Riemann, Joanna Schug, Brian Simpson, Caroline Ng Tseung-Wong and Wendy Johnson

Pers Soc Psychol Bull 2012 38: 1423 originally published online 27 June 2012
DOI: 10.1177/0146167212451275

The Effect of Response Style on Self-Reported Conscientiousness Across 20 Countries

Personality and Social Psychology Bulletin
38(11) 1423–1436

DOI: 10.1177/0146167212451275
http://pspb.sagepub.com

René Mõttus[1,2], Jüri Allik[1,3], Anu Realo[1], Jérôme Rossier[4], Gregory Zecca[4], Jennifer Ah-Kion[5], Dénis Amoussou-Yéyé[6], Martin Bäckström[7], Rasa Barkauskiene[8], Oumar Barry[9], Uma Bhowon[5], Fredrik Björklund[7], Aleksandra Bochaver[10], Konstantin Bochaver[10], Gideon de Bruin[11], Helena F. Cabrera[12], Sylvia Xiaohua Chen[13], A. Timothy Church[14], Daouda Dougoumalé Cissé[15], Donatien Dahourou[16], Xiaohang Feng[17], Yanjun Guan[18], Hyi-Sung Hwang[19], Fazilah Idris[20], Marcia S. Katigbak[14], Peter Kuppens[21,22], Anna Kwiatkowska[23], Alfredas Laurinavicius[24], Khairul Anwar Mastor[20], David Matsumoto[19], Rainer Riemann[25], Joanna Schug[26], Brian Simpson[19], Caroline Ng Tseung-Wong[5], and Wendy Johnson[2]

Abstract

Rankings of countries on mean levels of self-reported Conscientiousness continue to puzzle researchers. Based on the hypothesis that cross-cultural differences in the tendency to prefer extreme response categories of ordinal rating scales over moderate categories can influence the comparability of self-reports, this study investigated possible effects of response style on the mean levels of self-reported Conscientiousness in 22 samples from 20 countries. Extreme and neutral responding were estimated based on respondents' ratings of 30 hypothetical people described in short vignettes. In the vignette ratings, clear cross-sample differences in extreme and neutral responding emerged. These responding style differences were correlated with mean self-reported Conscientiousness scores. Correcting self-reports for extreme and neutral responding changed sample rankings of Conscientiousness, as well as the predictive validities of these rankings for external criteria. The findings suggest that the puzzling country rankings of self-reported Conscientiousness may to some extent result from differences in response styles.

Keywords

response style, extreme responding, Conscientiousness, cross-cultural, personality

Received October 8, 2011; revision accepted April 13, 2012

Introduction

In addition to comparing individuals within cultures, people's personality trait levels are often compared across cultures (e.g., McCrae & Terracciano, 2005; Schmitt, Allik, McCrae, & Benet-Martinez, 2007). The latter type of comparisons, especially, may sometimes lead to surprising and puzzling results. A good example is Conscientiousness, a broad Five-Factor Model (McCrae & John, 1992) personality trait that encompasses more specific traits such as being orderly, virtuous, traditional, self-controlled, responsible, and industrious (Roberts, Chernyshenko, Stark, & Goldberg, 2005). Rankings of countries (which are typically equated with cultures) on mean self-reported Conscientiousness scores (e.g., Schmitt et al., 2007) are often considered counterintuitive (e.g., Heine, Buchtel, & Norenzayan, 2008). The reason for this is that richer countries with higher life-expectancies (LEs) have generally *lower* mean scores of self-reported Conscientiousness than poorer countries with lower LEs; that is, the culture-level (often called ecological) correlations have been exactly opposite to the typical individual-level findings (Heine et al., 2008; but see also

[1]University of Tartu, Tartu, Estonia
[2]University of Edinburgh, Edinburgh, UK
[3]Estonian Academy of Sciences, Tallinn, Estonia
[4]University of Lausanne, Lausanne, Switzerland
[5]University of Mauritius, Réduit, Mauritius
[6]University of Abomey Calavi, Cotonou, Benin
[7]Lund University, Lund, Sweden
[8]Vilnius University, Vilnius, Lithuania
[9]University of Cheikh Anta Diop, Dakar, Senegal
[10]Moscow State University of Psychology and Education, Moscow, Russia
[11]University of Johannesburg, Johannesburg, South Africa
[12]University of Santo Tomas, Manila, Philippines
[13]Hong Kong Polytechnic University, Hong Kong, China
[14]Washington State University, Pullman, USA
[15]University of Bamako, Bamako, Mali
[16]University of Ouagadougou, Ouagadougou, Burkina Faso
[17]Changchun Normal University, , Chnagchun, China
[18]Renmin University of China, Beijing
[19]San Francisco State University, San Fransisco, USA
[20]Universiti Kebangsaan Malaysia, Bangi, Malaysia
[21]University of Melbourne, Melbourne, Australia
[22]University of Leuven, Leuven, Belgium
[23]Polish Academy of Sciences, Warsaw, Poland
[24]Mykolas Romeris University, Vilnius, Lithuania
[25]University of Bielefeld, Bielefeld, Germany
[26]Hokkaido University, Sapporo, Japan

Corresponding Author:
René Mõttus, University of Tartu, Tiigi 78, 50410 Tartu Estonia
Email: rene.mottus@ut.ee

Mõttus, Allik, & Realo, 2010; Mõttus et al., 2012). Of course, the ecological correlations do not necessarily have to mirror individual-level associations—in fact, they may be even completely opposite (Robinson, 1950)—but there is yet no good explanation as to why *high* mean levels of Conscientiousness should be associated with poverty and low average LE, which suggests the possibility that country-level mean Conscientiousness scores may be biased in some ways (Mõttus et al., 2012).

One potential source of bias is believed to be the reference group effect (RGE; Heine et al., 2008; Heine, Lehman, Peng, & Greenholtz, 2002), according to which different cultures have different subjective standards for traits. When people from different cultural settings base their subjective trait ratings on different standards, their ratings are incomparable and cross-cultural comparisons are therefore distorted. However, Mõttus and colleagues (2012) found only modest support for the RGE in Conscientiousness ratings across 22 samples representing different geographical and cultural groups. As a result, other potential biases should also be considered.

Another possible threat to cross-cultural comparisons of ratings based on ordinal rating scales (e.g., Likert-type or bipolar scales) comes from response styles which have been defined as systematic and pervasive tendencies "to respond to questionnaire items on some basis other than the specific item content" (Paulhus, 1991, p. 17). One such response style is extreme responding: the preference of extreme responses over more moderate ones when answering questionnaire items (Paulhus, 1991). Extreme response style may increase the variance of scores because people who tend to prefer more extreme responses will obtain higher (or lower) scores than those who choose moderate response categories, even when their true trait levels are identical. As a result, response style differences alone could potentially contribute to cross-country differences in the variability of test scores—a phenomenon often observed in personality traits, for example (Schmitt et al., 2007). Perhaps even more important is that extreme response style can also confound the comparisons of mean scores, which is one of the most essential methods of cross-cultural research (e.g., McCrae & Terracciano, 2005). If mean scores of self-report items systematically differ from the scale midpoints upward or downward, extreme response style will either inflate or depress these mean scores, respectively (Baumgartner & Steenkamp, 2001). This is because the chances that extreme responding inflates or deflates the scores are unequal if most people endorse the same side of the scale: In this case, extreme responding systematically produces either higher (if most people tend to endorse the higher end of the scale) or lower item scores (if the lower end of the scale is more often endorsed). Moreover, although extreme response style may often be associated with increased scale variance as said above, when item mean scores systematically differ from the scale midpoint, extreme response style may, in fact, decrease rather than increase the variance of the scores.

Having most or even all items of a multi-item trait measure systematically skewed in one direction—which would allow response style differences to create a systematic bias in mean trait scores—is not an unlikely scenario as people are generally known for the tendency to err on the socially desirable sides of rating scales (Krueger, 1998). Consistent with this, Mõttus and colleagues (2012) reported that in 20 of their 22 samples, mean scores of *all* self-reported Conscientiousness items were above the scale midpoint (i.e., toward higher levels of the trait), due to socially desirable responding or poor scale design, for example. Consequently, if there were differences across the samples in the preference for extreme responses to these Conscientiousness items, this may have distorted cross-sample comparisons of the self-reported trait scores.

There already exists a body of evidence demonstrating cross-cultural differences in the tendency to prefer extreme response categories of the ordinal rating scales over more moderate responses (Harzing, 2006). To give some examples, van Herk, Poortinga, and Verhallen (2004) investigated differences in response styles across six European countries and found that Greeks were the most likely to give extreme responses. Chen, Lee, and Stevenson (1995) reported that American students were more likely to choose extreme response categories of ordinal rating scales than Japanese and Chinese students, suggesting that Asian cultures may prefer a more moderate response style than Western cultures. Less is known about the response styles of African cultures though.

There is a difficulty, however, related to quantifying response styles such as extreme responding on the basis of self-report (or peer report) measures. The problem is that typically the phenomenon being rated (e.g., a personality trait or a value dimension) itself is expected to vary across the targets of the ratings or, at least, differences in how the phenomenon is perceived are likely to tap substantive variance (e.g., perception of national stereotypes). As a result, variance in the ratings simultaneously reflects at least two components: substantive variance due to veridical individual or cultural differences and variance due to biases such as extreme response style (Hamamura, Heine, & Paulhus, 2008). Therefore, attempts to estimate extreme response style on the basis of self-reports or peer reports (e.g., by calculating the ratio of extreme responses to more moderate responses) risk misinterpreting substantive variance as bias. For instance, in some countries people may indeed have higher levels of and/or vary more on personality traits compared with people in other countries (Schmitt et al., 2007), which, then, inclines them to gravitate toward the extreme response categories of test items.

The risk of confounding substantive variance and response bias is lower if extreme responding is estimated on the basis of items that measure different constructs and are uncorrelated (i.e., uncorrelated items reflect no single substantive construct to be confounded with response style; Greenleaf, 1992; Hamamura et al., 2008). Indeed, using uncorrelated items is a viable approach that can often be used to quantify

extreme response style end estimate its effect on cross-cultural comparisons. A particular strength of this method is that it does not require administering additional items. But this approach also has some potential downsides. First, it assumes that several unrelated constructs are measured at the same time because otherwise there will be no uncorrelated items available. As such, this is not a big problem because most surveys are likely to measure several constructs. Often, however, finding a sufficient number of uncorrelated items may be difficult even when multiple constructs, such as Big Five personality traits, are measured because the traits tend to be intercorrelated (e.g., Costa & McCrae, 1992; van der Linder, te Nijenhuis, & Bakker, 2010), for one reason or another. Second, and perhaps relatedly, response styles themselves can cause spurious intercorrelations between items (Baumgartner & Steenkamp, 2001), which may also make identifying a sufficient number of uncorrelated items more difficult.

It has also been suggested that response styles can be dealt with by ipsatizing scores (Fischer, 2004). This procedure standardizes respondents' (or groups') scores on a set of constructs in relation to their grand mean, so that respondents' (groups') scores on every trait become relative to the other scores of the same respondent (group). This, too, is a viable method but has some potential caveats. For example, this procedure also requires a number of constructs being measured at the same time. More important, however, is that ipsatization may change the substantive meaning of the transformed scores. For example, the grand mean differences between people or groups, which are removed with this procedure, may also convey meaningful information (besides possibly reflecting biases). In addition, this procedure does not allow a straightforward quantification of extreme response style, so its actual contribution to the scores remains difficult to estimate. For some researchers, having a direct and intuitively clear measure of response styles may be appealing, especially given that response styles may constitute interesting variables in their own right.

Taken together, additional and complementary ways of quantifying extreme response style will potentially be helpful for identifying and overcoming its possible confounding role in cross-cultural comparisons.

The Present Study

A method called "anchoring vignettes" could offer another possible solution for the above-described problem of possibly confounded true variance and response style effects. The method was originally developed in political sciences (King, Murray, Salomon, & Tandon, 2004; King & Wand, 2007) but was also used by Mõttus and colleagues (2012) to investigate the role of the RGE in cross-cultural comparisons of Conscientiousness ratings. In the study by Mõttus and colleagues (2012), nearly 3,000 people from 20 countries (22 samples in total) rated their own Conscientiousness and that of 30 hypothetical people described in short vignettes. These 30 hypothetical people portrayed six different facets of Conscientiousness from very low to very high levels of manifestation. The crucial feature of the study was that the vignettes were identical for all respondents. As a result, veridical individual differences among the rating targets, inherently present in self-ratings, were eliminated as a source of variance. Besides this advantage, two other factors made the vignette ratings helpful in teasing out systematic biases such as extreme responding from other sources of variance. First, because the vignettes were designed to display very different levels of Conscientiousness, the chances that all sorts of response categories would be widely chosen were increased and, as a result, there was ample room for individual differences in response styles to emerge. Second, with a relatively large number of targets to be rated, it was less likely that among-rater variance in aggregate estimates of response style reflected random measurement error; rather, it was likely that pervasive individual differences in rating biases such as extreme response style would ultimately "shine through."

Mõttus and colleagues (2012) found that sample-level mean vignette ratings were not consistently correlated with mean self-ratings, which offered little support for there being an RGE in cross-cultural comparisons of Conscientiousness. Reanalyzing the same unique data from a different perspective, the present study had three aims. First, it investigated whether there were differences across people from a wide range of geographical locations (22 samples in 20 countries) in the preference for extreme response categories of bipolar rating scales over moderate ones when rating Conscientiousness of the 30 hypothetical people, regardless of item content. Second, it tested whether the sample rankings on extreme responding covaried with the rankings of self-reported Conscientiousness scores, suggesting that the latter may have been confounded by differences in response style. Third, it examined whether correcting the sample rankings of self-reported Conscientiousness for response style differences had any effect on the rankings and their predictive validities for external criteria. Although linked to the same dataset, this study was different from the one by Mõttus and colleagues 2012, which focused exclusively on identifying the RGE—the original target of the anchoring vignettes method. The RGE is independent of extreme response style in both concept and measurement consequences. The RGE is based on the content of particular items, whereas extreme response style is defined as being independent of item content. Likewise, the possible effects of the RGE and extreme response style on cross-cultural comparisons of self-reports are orthogonal: both can either inflate or deflate self-reported mean scores of particular samples in completely independent ways.

Method

Participants

Altogether, 2,965 people from 20 countries participated in the study. The Peoples' Republic of China was represented

Table 1. Correlations Between the Response Category Choice Frequencies From Two Independent Subsets of Vignettes (N = 2,961)

	B				
A	Extreme left (neg)	Moderate left (neg)	Neutral	Moderate right (pos)	Extreme right (pos)
Extreme left (pos)	.36 (.44; .60)	–.28 (–.31; –.48)	–.30 (–.25; –.38)	–.47 (–.29; –.51)	.59 (.41; .63)
Moderate left (pos)	–.39 (–.31; –.47)	.29 (.30; .42)	.22 (.04; .20)	.43 (.26; .45)	–.46 (–.27; –.49)
Neutral	–.24 (–.24; –.38)	.02* (.05; .20)	.43 (.36; .48)	.18 (.12; .22)	–.31 (–.26; –.40)
Moderate right (neg)	–.41 (–.28; –.51)	.38 (.26; .46)	.11 (.11; .22)	.39 (.35; .46)	–.40 (–.40; –.52)
Extreme right (neg)	.55 (.40; .63)	–.35 (–.27; –.49)	–.35 (–.26; –.40)	–.44 (–.40; –.51)	.47 (.48; .63)

Note: A = response category choice frequencies based on the first subset of vignettes (rated for Competence, Dutifulness, and Self-Discipline; higher levels of the traits were endorsed by choosing the left-side categories); B = response category choice frequencies based on the second subset of vignettes (rated for Order, Achievement Striving, and Deliberation; higher levels of the traits were endorsed by choosing the right-side categories); pos = positive, higher levels of Conscientiousness; neg = negative, lower levels of Conscientiousness. The response category choice frequencies were calculated on items as they appeared to respondents (five categories ranging from extreme left to extreme right). In brackets are 2.5th and 97.5th percentiles of the distributions of 1,000 correlations calculated between respective response category frequencies from random subsets of vignettes (15 vignettes in each).
*This correlation was not statistically significant, whereas all other correlations were significant at $p < .001$ (with no adjustment for multiple testing).

with three independent samples—from Beijing, Changchun, and Hong Kong. Due to its high degree of autonomy and differing recent history, Hong Kong was treated as a separate sample. Also, because the other two Chinese samples from different locations were tested with independently translated testing materials, they were treated separately in all analyses. The 22 samples consisted exclusively of university students to keep the demographic profiles of the respondents similar. In the pooled sample, the mean age of participants was 22.17 years (SD = 5.27 years) and 62.56% of the participants were women. The demographic characteristics of the local samples are given in Table 1 of Mõttus and colleagues (2012).

Testing Materials

Conscientiousness was measured using six bipolar items that tapped six specific facets of the trait: Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, and Deliberation. The bipolar scales were taken from the National Character Survey (NCS; Terracciano et al., 2005). In the bipolar scales, the negative side of the trait was described on one and the positive side on the other end of the scale. Respondents were requested to mark one of the five different positions between the endpoints, with the middle one reflecting neutral or indecisive choice (i.e., not agreeing with either of the descriptions), the two response categories next to the middle category reflecting moderate preference for one of the endpoints, and the remaining two response categories reflecting extreme preference for one of the two endpoints. For instance, for the Deliberation facet, participants had to mark one of the five positions between the endpoints of the trait defined as "spontaneous, careless, thoughtless" and "cautious, reflective, careful." For Competence, Dutifulness, and Self-Discipline items, the descriptions reflecting high levels of the respective trait were on the left side; for the rest of the items they were on the right side: such item keying was retained throughout the analyses for the vignettes. However, self-reported Competence, Dutifulness, and Self-Discipline items were reversed before averaging all six self-report item scores to get a composite Conscientiousness score.

Five short descriptions of hypothetical people (vignettes) displaying various levels of the traits were drafted for each Conscientiousness facet (all vignettes are given in Appendix I of Mõttus et al., 2012). The five vignettes were intended to display different levels of the trait, from very low to very high. As for the NCS, the vignettes were first written in English and were then translated into the local languages where necessary, with the aim of retaining the meaning of the content as invariant as possible (except for the names of the people described in the vignettes: these were changed to better reflect local cultural circumstances). To ensure invariance in meaning, the translated testing materials were independently back translated into English and the back translations were reviewed by the first three authors. Where necessary, subsequent modifications of the translations were requested.

First, all participants rated their own Conscientiousness using the six bipolar scales. They then rated all 30 hypothetical people described in the vignettes using the same bipolar scales. Finally, respondents provided information about their ages and sex.

Four people chose the neutral (middle point "3") response category for 26 to 30 vignettes out of the 30; assuming that this reflected careless responding, the ratings from these respondents were excluded from further analyses. There was no evidence of such excessive use of other response categories.

External Criteria

As did Mõttus and colleagues (2012) and other studies (Heine et al., 2008; Mõttus et al., 2010; Oishi & Roth, 2009),

we used the per capita Gross Domestic Product (GDP) at Purchasing Power Parity in U.S. dollars and average LEs as objective sample-level criterion variables for mean self-reported Conscientiousness. The GDP and LE values were obtained from the Human Development Report (2009).

Results

Quantifying Response Styles

First, we tested whether people were consistent in their preferences for extreme response categories over more moderate ones, regardless of whether they rated low or high levels of Conscientiousness; without this, there would be little reason to talk about a pervasive extreme response style. Because different response choices were not independent from each other (choosing one response category automatically precludes choosing any other category), testing for possible consistencies in response scale use would have been difficult on the basis of a single set of vignettes. We therefore divided the 30 vignettes into two independent subsets and estimated consistencies across these two subsets. One subset (A in Table 1) included the vignettes that displayed various levels of and were rated for Competence, Dutifulness, and Self-Discipline (i.e., the vignettes that were rated with the bipolar scales depicting higher trait levels on the *left* side). The other subset (B in Table 1) included the remaining 15 vignettes that displayed various levels of and were rated for Order, Achievement Striving, and Deliberation (with high levels of the traits depicted on the *right* side of the bipolar scales). As mentioned above, such keying (for some scales high trait levels on the left and for others on the right) was retained throughout the analyses of the vignette ratings.

Separately for both subsets, we then calculated how frequently each respondent had used any of the five response categories to rate the 15 vignettes. To account for occasional missing responses, we divided the individual frequencies by the total number of responses given by the respondent. The five response categories, from left to right, had the following average frequencies across all respondents: 0.34, 0.12, 0.11, 0.16, and 0.27 for the first set and 0.22, 0.22, 0.14, 0.11, and 0.31 for the second set of vignettes. Then, the correlations between the response category choice frequencies in these two independent subsets were calculated at the level of the whole sample (Table 1). In addition, to test whether this particular division of vignettes produced different correlations than any other division, we randomly split the vignettes into two equally sized subsets (i.e., 15 vignettes in both) and calculated the correlations between the respective response category frequencies observed in the two random subsets. We repeated this procedure 1,000 times and report the 2.5th and 97.5th percentiles of the resulting correlations in Table 1.

There are four principal things to note in Table 1. First, those respondents who chose any of the five response categories more frequently than their peers in one subset also tended to prefer the same categories in the other subset (see the diagonal from top left to bottom right in Table 1). Similar patterns were generally observed when the vignettes were randomly split into subsets. These correlations also show that people who tended to give extreme negative (low Conscientiousness) ratings also tended to give extreme positive (high Conscientiousness) ratings, suggesting that the tendency to choose extreme responses was not restricted to either negative or positive trait levels. Second, despite the generally robust consistencies in preferring extreme or moderate response categories regardless of the pole of the dimension, there was some evidence for a slight tendency to prefer more strongly either negative or positive ratings (i.e., the valence also played a small role on top of the level of extremity). The correlations in the diagonal from top right to bottom left of the Table 1 (where the correlations address the frequencies of response categories that were exclusively *matched* in terms of valence [either high or low Conscientiousness] in addition to the level of extremity) were somewhat larger than those in the other diagonal (where the correlations compared the frequencies of response categories that exclusively *contradicted* in terms of valence [high vs. low Conscientiousness]). In addition, compared with the correlations that were calculated exclusively between extreme response categories that reflected opposing trait levels (correlations at top left and bottom right cells in the Table 1), the respective correlations from the randomly chosen subsets of vignettes were generally slightly higher.

Third, by necessity, the observed consistencies in preferring response categories meant that people who more often used extreme responses in either of the two subsets less often used moderate response categories in the other subset. This pattern, too, was confirmed in the random splitting procedure. Fourth, the consistencies in the response category frequencies appeared to depend somewhat on the particular subsets of vignettes the frequencies were based on. That is, there was some variability in the correlations obtained from random subsets of vignettes, and in a few cases, these correlations did not overlap with those from the original subsets A and B. Despite this, the overall pattern was robust: no matter which subsets were used, individual differences in the tendencies to prefer either extreme or modest responses were always evident. Of note is that a similar pattern of associations was observed at the level of the mean frequencies of the 22 samples: higher average use of any of the five response categories in one subset of vignettes was associated with higher average use of the same response categories in the other subset.

Put simply, the detailed evidence presented above indicates that people were relatively consistent in their preferences for extreme over moderate responses, largely regardless of the side of the scale involved. Therefore, the frequencies of extreme responses on both sides of rating scales could be aggregated for the whole set of 30 vignettes and so could the frequencies of moderate responses.

Next, we calculated the proportion of extreme to total responses, excluding neutral middle-point response category choices, to form an index of extreme responding (ER) for each respondent. Designating the five response categories (from left to right) "1," "2," "3," "4," and "5," the ER was thus calculated as follows: ER = ("1" + "5")/("1" + "2" + "4" + "5"). The proportion of neutral middle-point responses was not included in calculating the ER because, unlike the other response categories, it did not reflect any sort of agreement with the descriptions at scale endpoints. Instead, it reflected neutrality or indecisiveness. Therefore, the ratio of neutral middle-point responses to the total number of responses was used as a separate, and complementary, index of neutral responding (NR). The two indices, ER and NR, were necessarily negatively correlated at $r = -.33$ (for the whole sample; $p < .001$). Because the study mainly focused on cross-sample differences, the means of the ER and NR for the 22 samples are given in Table 2.

Response Profiles: The "Content" of the Response Style Indices

Figure 1 illustrates the "content" of the sample mean ER and NR, displaying the sample mean frequencies for each of the five response categories chosen to rate the 30 vignettes. For visual clarity, mean frequencies are given only for three samples: the sample with the lowest ER (Hong Kong, China), a sample with a close-to-median ER (Estonia), and the sample with the highest ER (Changchun, China). The profile for Estonia was fairly similar to what would have been the average profile across all samples. With only five response categories, it is easy to see how the two indices effectively summarize the whole distribution of response frequencies. The ER quantifies the proportion of extreme agreement to total (both extreme and moderate) agreement, that is, ER simultaneously summarizes the "slopes" at both sides that are the steepest for Changchun (China) sample and almost flat for the Hong Kong (China) sample. Because the sum of the frequencies of the five response categories is fixed (i.e., the total number of responses), the NR basically summarizes the rest: the degree of neutrality of responding (or disagreement with either of the scale ends) and, at the same time, the "intercept" for the ER "slope." Because the two statistics treated the middle-point response category differently, the NR rankings of the samples were different from their rankings of ER, with Estonians using neutral responses the least of the three sample and Hong Kong respondents using them the most among the three.

Cross-Sample Differences in Response Style

Using general linear models, we next estimated the amounts of variance in the two indices attributable to differences among the samples. Because differences among the samples in mean ages and the proportions of women (Mõttus et al., 2012; Table 1) could have confounded the effects of sample on the ER and NR per se, age and sex were included as covariates in the models. A fair amount of the variance in the ER was accounted for by among-sample differences (8.2%, partial eta-squared [η^2_p] = .082, $p < .001$). Sex explained 2% of variance ($\eta^2_p = .020$, $p < .001$, women had higher ERs), whereas the age effect was not statistically significant. For the NR, 5.7% of variance was accounted for by among-sample differences ($\eta^2_p = .057$, $p < .001$) and far less by sex ($\eta^2_p = .007$, $p < .001$, men higher); the age effect was not significant. For comparison purposes, in self-rated Conscientiousness (mean of the six bipolar items for self), differences among samples, age, and sex explained 15.3% ($p < .001$), 0.6% ($p < .001$), and 0.5% ($p < .001$) of total variance, respectively (older respondents and women had higher scores). In sum, there was a clear pattern of cross-sample differences in response styles as quantified on the basis of the vignette ratings. The lowest rates of extreme responding characterized Hong Kong (China), South-Korea, Germany, and Japan, whereas several African (e.g. Benin, South Africa, Senegal, Burkina Faso) and Southeast Asian (Malaysia, Philippines) samples, as well as Polish and Changchun (China) samples, had the highest rates (Table 2). Most European nations, Australia, and the United States were characterized by medium rates of extreme responding.

Generalizability of the Response Styles From Vignettes to Self-Reports

The response styles (ER and NR) in the vignette ratings generalized to some extent to those calculated on the six self-report items. In the latter, the response patterns were likely to characterize true individual differences in addition to response styles, and we might expect population variance in Conscientiousness to be smaller than that in the vignettes as these were designed to display very different levels of the trait; therefore, the correlations were not expected to be strong. For the whole sample, the ER calculated on the six self-report items was correlated with the ER obtained from the 30 vignette ratings at $r = .26$ ($p < .001$). The correlation between NRs based on self-reports and vignette ratings was somewhat lower ($r = .15$, $p < .001$). Our primary focus, however, was on the associations occurring at the level of sample means and not at the level of single individuals, as is typical in studies that compare countries. At the level of sample means (i.e., ignoring all within-sample variance), the ecological associations may appear very different from the associations occurring at the level of individuals, where between-sample differences are mixed with other sources of variance. As is sometimes the case, at the level of sample means, the correlations between response style indices based on vignettes and self-reports were higher, with the Spearman correlations being rho = .71 ($p < .001$) and rho = .35 ($p = .11$) for ER and NR, respectively. Therefore, there was a strong tendency for sample-level extreme response style in the

Table 2. Mean Response Style Indices Based on Vignette Ratings for the 22 Samples

	ER (*M*)	ER (*SD*)	ER (rankings)	NR (*M*)	NR (*SD*)	NR (rankings)
Hong Kong (China)	0.49	0.21	1	0.14	0.09	19
South-Korea	0.56	0.17	2	0.14	0.08	14
Germany	0.58	0.14	3	0.14	0.07	16
Japan	0.59	0.20	4	0.13	0.09	11
Mauritius	0.60	0.20	5	0.15	0.11	21
Sweden	0.63	0.17	6	0.15	0.07	22
Australia	0.64	0.16	7	0.09	0.07	2
Beijing (China)	0.64	0.17	8	0.11	0.07	5
Lithuania	0.64	0.16	9	0.13	0.08	12
Switzerland	0.64	0.14	10	0.14	0.08	15
Estonia	0.65	0.15	11	0.12	0.07	9
Mali	0.65	0.17	12	0.09	0.09	1
Russia	0.65	0.15	13	0.14	0.08	20
USA	0.67	0.15	14	0.14	0.10	17
Benin	0.68	0.27	15	0.12	0.08	8
South Africa	0.68	0.21	16	0.14	0.09	18
Malaysia	0.69	0.17	17	0.11	0.08	3
Senegal	0.69	0.22	18	0.11	0.08	4
Philippines	0.70	0.18	19	0.12	0.07	7
Poland	0.70	0.17	20	0.12	0.09	6
Burkina Faso	0.71	0.20	21	0.12	0.08	10
Changchun (China)	0.72	0.18	22	0.14	0.09	13
Grand *M*	0.64			0.12		
Grand *SD*	0.18			0.08		

Note: ER = average index of extreme responding of the samples; NR = average index of neutral responding of the samples.

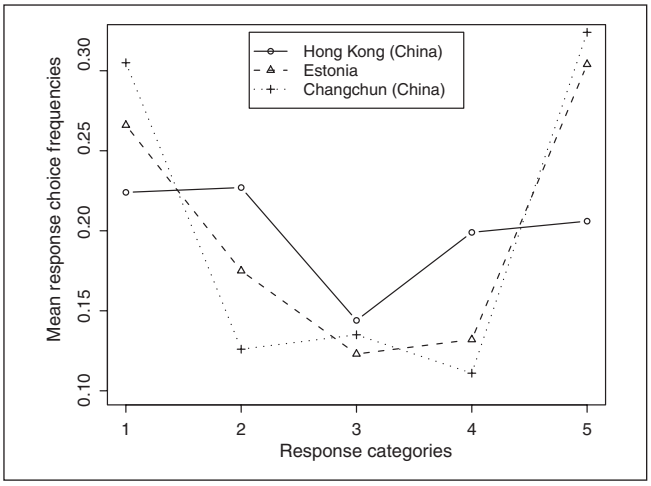

Figure 1. Mean response category choice frequencies in three samples
Note: Category numbers 1 to 5 designate the five response categories of the bipolar scale from extreme left to extreme right as they appeared to respondents.

vignette ratings to track with sample-level extreme response style in self-reports. Here and hereafter, Spearman rank-order correlation (rho) was used for analyzing sample-level associations because with only 22 values, Pearson correlations are especially sensitive to any violations of normality (e.g., due to outliers such as Hong Kong [China] for ER and Japan for mean self-reported Conscientiousness, see Tables 2 and 3).

The Associations of Response Styles With Self-Reported Conscientiousness Scores

In nearly all samples, means of the six self-reported Conscientiousness items deviated from the scale midpoint (3) in the direction of higher Conscientiousness (Mõttus et al., 2012; Table 2), which created a possibility for inflation of mean self-reports due to extreme responding (Baumgartner & Steenkamp, 2001). Assessment of the correlations between response styles and self-reported Conscientiousness provided a test of this possibility. At the level of the whole sample, self-reported Conscientiousness (mean of the six bipolar items) was positively correlated with the vignette-based ER ($r = .20$, $p < .001$). Excluding the within-samples variance by estimating the association at the level of sample means, the correlation was much higher (rho = .70, $p < .001$). That is, in the samples with the lowest mean self-reported Conscientiousness (Table 3), South-Korea and Japan, people also tended, on average, to use extreme response categories relatively less frequently to rate the 30 vignettes than people

Table 3. Mean Self-Reported Conscientiousness for the 22 Samples Before and After Correcting for Response Styles

	C (*M*)	C (*SD*)	C (rankings)	C_{res}	C_{res} (rankings)
Japan	3.13	0.77	1	−1.21	4
South-Korea	3.45	0.55	2	−0.65	9
Lithuania	3.54	0.53	3	−1.30	2
Australia	3.62	0.64	4	−1.21	3
Russia	3.67	0.66	5	−1.56	1
Switzerland	3.69	0.52	6	−0.90	7
Estonia	3.73	0.58	7	−0.92	6
Mauritius	3.74	0.62	8	0.36	13
Hong Kong (China)	3.76	0.67	9	1.16	20
Germany	3.77	0.59	10	1.02	18
Sweden	3.85	0.50	11	0.90	16
Malaysia	3.89	0.56	12	−0.99	5
USA	3.93	0.54	13	0.06	12
Poland	3.94	0.65	14	−0.66	8
Beijing (China)	3.96	0.58	15	1.05	19
Philippines	3.97	0.56	16	−0.35	10
South Africa	4.01	0.58	17	0.66	14
Changchun (China)	4.06	0.62	18	−0.12	11
Mali	4.23	0.43	19	1.47	21
Senegal	4.25	0.53	20	0.93	17
Burkina Faso	4.27	0.55	21	0.66	15
Benin	4.37	0.53	22	1.69	22
Grand *M*	3.83				
Grand *SD*	0.65				

Note: C = Conscientiousness; C_{res} = standardized regression residuals after regressing sample mean Conscientiousness scores (column 2) on the mean extreme and neutral responding scores (given in Table 2).

from most of the other samples. In contrast, respondents from Burkina Faso, Senegal, and Benin had the highest mean self-reported Conscientiousness scores, and they also tended to be among the frequent users of extreme response categories while rating the vignettes. The vignette-based NR was negatively but less strongly correlated with self-reported Conscientiousness (rho = −.08, $p < .001$, for individual respondents and rho = −.32, $p = .14$, at the level of sample means). We repeated the analyses for single Conscientiousness facets and obtained a largely similar pattern of results. At the level of individual respondents, correlations between ER (correlations for NR in parentheses) and Conscientiousness facets ranged from $r = .09$ to .18 with a median of .13 (from $r = -.03$ to −.07 with a median of −.05). At the level of sample means, correlations between ER (NR) and Conscientiousness facets ranged from rho = .42 to .74 with a median of .57 (from rho = −.03 to −.52 with a median of −.28).

Correcting Self-Reports for Response Styles

Next, assuming that response styles may potentially have affected self-report scores, we tested to what degree removing the effects of extreme and neutral responding from self-reported Conscientiousness could change the rankings of people and samples on this trait. To this end, Conscientiousness scores were residualized (using multiple regression) for the vignette-based ER and NR, first at the level of individual respondents and then at the level of sample means. For individual respondents, the residualized scores correlated highly with the original Conscientiousness scores ($r = .98$, $p < .001$), suggesting essentially no effect of individual response styles on individual self-reports in the sample as a whole. However, because at the level of sample means the correlations between response style indices and self-reported Conscientiousness had been much higher (although nonsignificant for the NR due to the small number of samples), the effect of residualizing sample-level mean Conscientiousness scores for the mean response style indices was expected to be stronger. Before performing the regression to obtain residualized sample-level Conscientiousness scores, sample means of all variables were transformed to rankings to avoid the confounding effect of nonnormality (e.g., Hong Kong's outstandingly low mean ER value and Japan's outstandingly low mean Conscientiousness score). The hypothesis was confirmed as the correlation between the original and residualized sample

rankings of Conscientiousness was lower (rho = .68, $p < .001$) than the individual-level correlation of .98 given above. Repeating the analyses for single Conscientiousness facets yielded generally similar results. At the level of individual respondents, the correlations between original and residualized facet scores ranged from $r = .98$ to 1.00. At the level of sample means, correcting for response styles had the strongest effects for Dutifulness and Competence (rho = .59 and .67, $p < .01$, respectively) and weakest effects for Order and Deliberation (rho = .91 and .87, $p < .001$, respectively); the median correlation between the original and the residualized sample rankings was rho = .81.

The rankings of samples based on both the original and residualized mean Conscientiousness scores are given in Table 3. Among the biggest changes, Hong Kong (China) moved 11, Germany 8, South-Korea 7, Mauritius and Sweden 5, Beijing (China) 4, and Japan 3 positions upward, whereas Malaysia and Changchun (China) moved 7, Burkina Faso, Philippines, and Poland 6, Russia 4, and Senegal and South Africa 3 positions downward in terms of mean Conscientiousness scores.

The Effect of Response Styles on the Predictive Validity of Conscientiousness Rankings

As correcting for response style differences had a notable effect on sample rankings on Conscientiousness, we tested whether the correction also reflected in the correlations of these rankings with GDP and LE. The original sample-level mean Conscientiousness scores correlated with these variables at rho = −.71 ($p < .001$) and rho = −.65 ($p < .01$), respectively. The residualized mean Conscientiousness scores, however, had notably lower correlations with GDP and LE, rho = −.33 ($p = 0.13$) and rho = −.26 ($p = 0.24$), respectively. We obtained similar findings for single Conscientiousness facets: Uncorrected sample means of the six facets correlated with GDP in the range of rho = −.39 to −.69 (median rho = −.65), whereas the residualized sample means had much lower correlations (from rho = .00 to −.47, median = −.28). For LE, the respective correlations ranged from rho = −.31 to −.68 (median rho = −0.57) and from rho = −0.15 to −0.54 (median rho = −.25).

Discussion

Cross-cultural comparative research is based on the assumption that measurements made in different cultural contexts (e.g., countries, regions) are comparable. That is, when trait levels are compared, the same observed levels should correspond to the same true trait levels in all groups under comparison. The most serious threat to this assumption comes from systematic biases in the observed trait levels. Therefore, identifying any possible systematic biases and developing means for overcoming them are essential for the development of cross-cultural comparative research.

Using a novel approach for separating response style effects on self-reports from the true variance of traits, the present study demonstrated cross-cultural differences in the tendency to prefer extreme response categories of bipolar items over more moderate ones when rating the personality trait Conscientiousness. Although there was generally little variance in extreme response style among most European, American, and Australian samples, respondents from other world regions often displayed different degrees of preference for the extreme responses. Both individual and cross-sample differences in the tendency to use extreme response categories as quantified on the basis of the vignette ratings were associated with extreme response style as observed in self-reports and—more important—with the self-reported Conscientiousness scores themselves. As is often (but certainly not always) the case with such ecological correlations, these associations were particularly strong at the aggregate level of the 22 samples (Spearman's rho = .70). Controlling for cross-sample differences in extreme response style (as well as the preference for the neutral middle-point response categories) had notable effects on the rankings of the samples on Conscientiousness. The corrections also changed the predictive validities of these rankings for GDP and average LE. Below, it will be discussed whether the changes in the predictive validities can be considered meaningful.

The present study focused on the identification and possible consequences of cross-cultural differences in response styles and not on explaining the observed response style differences. For the sake of completeness, however, a brief comment on the latter is warranted. A look at Table 2 readily shows that lower mean levels of extreme responding are associated with higher economic and social development and East Asian cultures, whereas high mean levels of extreme responding mainly (but not exclusively) characterize economically less developed countries and African and Southeast Asian cultures. Besides a very general explanation that higher levels of societal development (e.g., higher mean educational level) may incline people to, on average, abstain from overly extreme judgments, one might hypothesize that cross-cultural differences in what is called dialectical thinking may contribute to the variations in response styles. Dialectical thinking is characterized by "an emphasis on change, a recognition of contradiction and of the need for multiple perspectives, and a search for the 'Middle Way' between opposing propositions" (Nisbett, Peng, Choi, & Norenzayan, 2001, p. 293). As a result, dialectical thinking may lead to less extreme and polarized subjective judgments because low and high trait levels can easily coexist and change for dialectical thinkers. It has been hypothesized that East Asian cultures are characterized by higher degrees of dialectical thinking than Western or African cultures (e.g., Schimmack, Oishi, & Diener, 2002; Spencer-Rodgers, Williams, & Peng, 2010). There is indeed some empirical evidence for higher levels of dialectical

thinking being associated with less polarized judgments (Hamamura et al., 2008; Minkov, 2009).

However, the limited number of countries used in the present study and even more limited overlap with the existing datasets on dialectical thinking (for which especially little data are available for African cultures, for example, Schimmack et al., 2002) prevented us from formally testing these associations. We hope that future studies will continue to investigate the possible role of dialectical thinking on cross-cultural variability in response styles empirically, as it appears in self-ratings or ratings of other people (e.g., vignettes). Likewise, although differences in dialectical thinking seem currently one of the most plausible explanations for geographical differences in extreme or neutral response styles, future studies may consider other theoretically relevant constructs. However, for any explanations, it will be important to make sure that the scores on the explanatory variables themselves are not confounded by response styles (van Herk et al., 2004).

Theoretical Implications of the Findings

The results of this study suggest that the puzzling country rankings of Conscientiousness may, to some extent, result from cross-cultural differences in the tendencies to prefer extreme response categories of self-report measures over more moderate response categories. After adjusting for the response style differences, samples from Changchun (China), Malaysia, Burkina Faso, Philippines, and Poland that had high prevalences of extreme responding slipped downward in the rankings of mean Conscientiousness scores. In contrast, Hong Kong (China), Germany, South-Korea, Mauritius, Sweden, and Japan where respondents somewhat less often chose extreme response categories to rate the vignettes moved upward in mean Conscientiousness. After adjusting for the response style differences, the counterintuitive, as some authors think (Heine et al., 2008), correlations with GDP and average LE were also notably attenuated and were no longer statistically significant. Thus, although correcting for the response style differences certainly did not reverse the Conscientiousness rankings of samples and their correlations with external criteria, it had a clear effect.

Of course, although it is sometimes thought that *negative* correlations between mean Conscientiousness and national economic output or average LE demonstrate the *in*validity of mean Conscientiousness scores (Heine et al., 2008), alternative interpretations are also possible. It may be that the direction of the observed associations is in fact meaningfully interpretable (Hofstede & McCrae, 2004; Mõttus et al., 2010; Mõttus et al., 2012). However, what may be even more worrisome about the observed uncorrected associations is their strength. For example, in this study, mean Conscientiousness scores explained half of the variance in GDP (a very similar correlation was reported by Mõttus et al., [2010] on a larger number of countries and using another self-report measure of Conscientiousness). Taking into account possible unreliability of the Conscientiousness measure, this is a very strong association indeed. Although ecological correlations are often high, this is by no means inevitable or trivial (for a discussion, see Asendorpf's comment in Allik et al., 2007). Should this be causally interpreted as national differences in the lack of Conscientiousness accounting for more than half of differences in economic output? This may be highly unrealistic considering that there is probably a myriad of reasons why nations differ in their economic output in a given year. The converse is also true: Expecting national differences in economic output in a given year to cause the majority of the cross-country variance in personality scores is simply not realistic. Sometimes, thus, it is precisely the strength (not the weakness or absence) of the observed associations that is theoretically most alarming (Lykken, 1968). If this line of reasoning is true, this leaves us with the Conscientiousness–GDP associations being confounded on top of, or even instead of, any substantive associations. Therefore, the more modest, albeit nonsignificant due to a small number of samples, validity correlations after correcting self-reported Conscientiousness for response styles are perhaps in a more meaningful range than the uncorrected associations.

These results also suggest that response styles will contribute to difficulties in achieving full measurement invariance across a wide range of cultures when assessing Conscientiousness (and possibly other traits) by means of self-reports. Lack of measurement invariance means that trait scores obtained from different samples do not reflect exactly the same trait to the same degree, due to indicators defining the trait with different loadings, intercepts, and/or residual variances. It has been shown that differences in extreme responding affect both factor loadings and intercepts of observed indicator scores on latent personality traits (Cheung & Rensvold, 2000). There being contributions from cross-cultural differences in response styles to measurement noninvariance would be consistent with the existing reports describing difficulties in establishing measurement equivalence of personality traits across cultures (e.g., Church et al., 2011; Johnson, Spinath, Krueger, Angleitner, & Riemann, 2008; Rossier, Dahourou, & McCrae, 2005). Of course, it must be noted that response style differences may be only one source of cross-cultural measurement noninvariance of personality traits. However, if the effects of response style differences on mean self-reported Conscientiousness and other personality trait scores prove to be replicable and causal in future studies, their measurement invariance implications will need to be heeded in cross-cultural personality research.

Alternative Interpretations

Correcting the rankings of self-reported Conscientiousness for response style differences was based on the hypothesis that differences in response styles, as measured on the basis of the vignette ratings, could potentially contribute to the

observed differences in self-reported Conscientiousness. That is, we hypothesized that these were the response styles that may have distorted the rankings of self-reports rather than the other way around. However, we are fully aware that there are alternative ways to interpret the association between response styles and self-reported Conscientiousness (Austin, Deary, & Egan, 2006). For example, it is possible that the sample rankings of self-reported Conscientiousness were in fact accurate, and it was living in highly conscientious cultural settings (e.g., in Burkina Faso) that made respondents use extreme response categories rather than moderate response categories while rating the vignettes. Or, response styles and self-reported Conscientiousness may have covaried due to unknown common determinants. Thus, although it is easy to see how response styles can affect self-reported trait scores when most people, for whatever reason, prefer one side of Likert-type or bipolar rating scales (as was described above), there is no strict empirical evidence as yet for preferring this causal explanation over the alternative ones.

For an ultimate test of which explanation is most plausible, we would need to investigate the associations between response style indices and Conscientiousness *as measured independently of self-reports* (see McCrae & Costa, 1983). If response style indices were associated with self-reported Conscientiousness scores but not with the alternative and independent operationalizations of the trait, it would indicate that response styles are likely to be causal contributors, beyond actual trait levels, to self-reported trait scores. In contrast, if response styles were similarly associated with alternative operationalizations of Conscientiousness, it would probably mean that Conscientiousness itself determines response styles or that both result from some overlapping unknown causes. However, there are currently no good ways to measure cross-cultural differences in Conscientiousness independently of self-reports (or related methods). Note that even peer reports are not helpful here because the cross-cultural differences in response styles are likely to generalize to all types of ratings made using ordinal rating scales, so similar culture differences are likely also to appear in peer reports.

Therefore, as long as there is no empirical way of testing whether extreme responding indeed confounds the observed mean self-report scores or is simply a yet another manifestation of veridical cross-sample differences in Conscientiousness, we have to rely on common sense to interpret the association between extreme response style in vignette ratings and self-reported Conscientiousness. It is currently difficult to give a theoretical explanation for why higher mean Conscientiousness (the same high mean Conscientiousness that is very strongly predictive of low national wealth and low mean LE) should causally make people prefer extreme responses over moderate ones. If anything, the opposite could be expected because one of the Conscientiousness facets is Deliberation, which, in the present study, was defined as being cautious, reflective, and careful: It is perhaps commonsensical to expect cautious people to refrain from extreme statements such as giving extreme trait ratings when somewhat limited information about the targets is available. Likewise, we cannot think of any meaningful common determinants of both response styles and mean Conscientiousness scores. Therefore, it currently seems most reasonable to believe that an explanation which has all necessary elements in place (as has been explained above, we can see the "mechanics" of how extreme response style can affect self-reported Conscientiousness scores under the present circumstances) could be preferred to explanations that are possible but do not have any theoretical account as yet to support them.

Practical Implications

In addition to the substantive contribution, this study featured vignettes as potentially useful practical tools for identifying and mitigating extreme response style. Could this method be useful in future cross-cultural psychological research? Based on the fact that it is already being used in areas such as health (Grol-Prokopczyk, Freese, & Hauser, 2011) or economics (Kristensen & Johansson, 2008) surveys—with backing from the statistical community (van Soest, Delaney, Harmon, Kapteyn, & Smith, 2011)—to identify and overcome the RGE-type measurement issues, there is no fundamental reason, at least, why it could not be practically used in psychological research. Like all methods, it has both strengths and limitations, which make it more suitable for some research purposes than for others. Perhaps the most important strength of the approach is that it can be used for more than one purpose. Although in many cases alternative methods for quantifying response styles or dealing with their consequences are available—such as calculating extreme responding on the basis of a set of uncorrelated items (Greenleaf, 1992; Hamamura et al., 2008) or ipsatizing scores (Fischer, 2004)—the advantage of the anchoring vignette approach (King & Wand, 2007) is that it provides a more generic, yet simple and intuitive method for simultaneously detecting various types of biases, such as trait-specific RGE (Mõttus et al., 2012) or response styles that cut across constructs. The most important issue with the method is cost—it requires additional survey items (vignettes) to be administered.

The decisions about whether the strengths outweigh the cost or exactly how much needs to be invested in the vignette approach probably depend on what researchers are most worried about. If response styles are the only possible source of threat for the comparability of self-reports, then researchers may use other methods for detecting bias (Greenleaf, 1992; Hamamura et al., 2008). Alternatively, they may administer a limited number of vignettes: Because response styles are, by definition, independent of specific item content, response styles quantified on the basis of one trait are likely to generalize to other trait. In addition, there is probably no need to administer 30 vignettes to quantify response styles for one trait (Mõttus and colleagues [2012] administered 30 vignettes because they wanted to address six specific facets

of Conscientiousness having 5 vignettes for each facet). Perhaps five or even less vignettes for one or more questionnaire items can provide enough information to quantify response styles.

However, if researchers cannot rule out the existence of an RGE-type of bias, the vignette method could be used to its full potential. Then, vignettes should be administered for all of the traits that may potentially suffer from the biases. For example, researchers can choose one to three items from each domain (e.g., the Big Five domains) and administer three or more vignettes for each. Of note is that dealing with neither response styles nor RGE (King & Wand, 2007) strictly assumes that all respondents have to be administered the vignettes: The biases can be identified using only subsamples of each sample and then generalized to populations. Of course, in some cases, even this may be too costly, whereas in some cases the price of not fully addressing the problems may outweigh the cost of additional survey items. One of such cases where the price of not properly dealing with possible biases may be especially high is when researchers are faced with puzzling findings such as the country rankings of Conscientiousness.

Strengths and Limitations of the Study

The primary strengths of this study were the novel method for disentangling response bias from true variance and the ability to see whether the response style differences across people and samples measured using the ratings of invariant targets were associated with self-reports of the same people and samples. Also a noteworthy strength is the range of cultures incorporated; for instance, to date little was known about extreme response style in African samples.

Among the limitations is the not particularly large number of samples, which may influence the reliability of the sample-level estimates. Moreover, the results of all cross-cultural studies highly depend on the comparability of the translations of testing material and, despite the efforts that were made to grant equivalency of the measures, this study was no exception. However, the cross-sample differences in response style were probably not caused by differences in translations because in several samples identical translations were used but response styles differed. In particular, the Hong Kong and Beijing Chinese samples were tested with the same Chinese translation, Switzerland and several African samples (e.g., Senegal and Burkina Faso) were tested with the same French translation, and Australia, USA, and South Africa were tested with the same English translation; yet, response styles were different. Other confounding sources of cross-cultural variance, however, remain possible. Finally, different types of response scales (e.g., Likert-type) or scales with different numbers of response categories (e.g., 3, 7, 9) may have resulted in different results (Hui & Triandis, 1989).

Conclusion

The RGE has been the primary suspect for distorting cross-cultural comparisons of mean Conscientiousness scores (Heine et al., 2008). However, the first use of the anchoring vignettes method (King & Wand, 2007) in cross-cultural personality research provided only little evidence for RGE affecting country rankings of Conscientiousness (Mõttus et al., 2012). Extending the applicability of the method to a completely different source of bias, this study showed that response style, especially extreme responding, is a far stronger candidate for distorting country rankings of Conscientiousness than the RGE. Beyond the particular problem of geographical variations in Conscientiousness, the results of this study show that quantifying response styles on the basis of vignette ratings is likely to be helpful in identifying differences in response styles and, equally importantly, in correcting for their effects. What is more, the method allows researchers to identify different sources of measurement bias at the same time. Thus, the study made a unique substantive contribution to the literature in potentially moving toward an explanation of the paradox of mean Conscientiousness scores, but it also made a unique methodological contribution in extending the applicability of the anchoring vignettes approach to dealing with response style problems in cross-cultural measurement and beyond this.

Authors' Note

The data used in the study are available for reanalyses from the first author.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was supported by grants from the Estonian Ministry of Science and Education (SF0180029s08) and the Estonian Science Foundation (ESF7020) to Jüri Allik, by a Swiss National Science Foundation grant (ZK0Z1_131287/1) to Jüri Allik and Jérôme Rossier, by a Mobilitas grant (MJD44) from the European Social Fund to René Mõttus, and by a Primus grant (3-8.2/60) from the European Social Fund to Anu Realo.

References

Allik, J., Asendorpf, J. B., Bosker, R. J., Brunner, M., Martin, R., Ceci, S. J., Williams, W. M. . . . Wilhelm, O(2007). Discussion on "The g-Factor of International Cognitive Ability Comparisons: The Homogeneity of Results in PISA, TIMSS, PIRLS and IQ-Tests Across Nations" by Heiner Rindermann. *European Journal of Personality, 21*, 707-765.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40*, 1235-1245.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156.

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6*, 170-175.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*, 187-212.

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology, 101*, 1068-1089.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Fischer, R. 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.

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly, 56*, 328-351.

Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior, 52*, 246-261.

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences, 44*, 932-942.

Harzing, A.-W. (2006). Response styles in cross-national survey research. *International Journal of Cross Cultural Management, 6*, 243-266.

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science, 19*, 309-313.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology, 82*, 903-918.

Hofstede, G., & McCrae, R. R. (2004). Personality and culture revisited: Linking traits and dimensions of culture. *Cross-Cultural Research, 38*, 52-88.

Hui, C. H., & Triandis, C. H. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*, 296-309.

Human Development Report. (2009). *Human development index 2007*. New York, NY: The United Nations Development Program. Available from http://hdr.undp.org/en/reports/global/hdr2009/

Johnson, W., Spinath, F., Krueger, R. F., Angleitner, A., & Riemann, R. (2008). Personality in Germany and Minnesota: An IRT-based comparison of MPQ self-reports. *Journal of Personality, 76*, 665-706.

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.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis, 15*, 46-66.

Kristensen, N., & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics, 15*, 96-117.

Krueger, J. (1998). Enhancement bias in descriptions of self and others. *Personality and Social Psychology Bulletin, 24*, 505-516.

Lykken, D. T. (1968). Statistical significance of psychological research. *Psychological Bulletin, 70*, 151-159.

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*, 882-888.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215.

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

Minkov, M. (2009). Nations with more dialectical selves exhibit lower polarization in life quality judgments and social opinions. *Cross-Cultural Research, 43*, 230-250.

Mõttus, R., Allik, J., & Realo, A. (2010). An attempt to validate national mean scores of Conscientiousness: No necessarily paradoxical findings. *Journal of Research in Personality, 44*, 630-640.

Mõttus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., . . . Tseung, C. N. (2012). Comparability of self-reported Conscientiousness across 21 countries. *European Journal of Personality, 26*, 307-317. doi:10.1002/per.840

Nisbett, R. E., Peng, K. P., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*, 291-310.

Oishi, S., & Roth, D. P. (2009). The role of self-reports in culture and personality research: It is too early to give up on self-reports. *Journal of Research in Personality, 43*, 107-109.

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

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology, 58*, 103-139.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351-357.

Rossier, J., Dahourou, D., & McCrae, R. R. (2005). Structural and mean-level analyses of the five-factor model and locus of control further evidence from Africa. *Journal of Cross-Cultural Psychology, 36*, 227-246.

Schimmack, U., Oishi, S., & Diener, E. (2002). Cultural influences on the relation between pleasant emotions and unpleasant emotions: Asian dialectic philosophies or individualism-collectivism? *Cognition & Emotion, 16*, 705-719.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martinez, V. (2007). The geographic distribution of big five personality traits—Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*, 173-212.

Spencer-Rodgers, J., Williams, M. J., & Peng, K. (2010). Cultural differences in expectations of change and tolerance for contradiction: A decade of empirical research. *Personality and Social Psychology Bulletin, 14*, 296-312.

Terracciano, A., Abdel-Khalek, A. M., Adam, N., Adamovova, L., Ahn, C., Ahn, H. N., . . . McCrae, R. R. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310*, 96-100.

van der Linder, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*, 315-327.

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales. *Journal of Cross-Cultural Psychology, 35*, 346 –360.

van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*, 575-595.