Comparability of Self-Reported Conscientiousness Across 21 Countries

European Journal of Personality

Comparability of Self-Reported Conscientiousness across 21 Countries

Journal:	*European Journal of Personality*
Manuscript ID:	EJP-11-1052.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	18-Apr-2011
Complete List of Authors:	Mõttus, Rene; University of Tartu, Psychology Allik, Jüri; University of Tartu, Department of Psychology Realo, Anu; University of Tartu, Department of Psychology Pullmann, Helle; University of Tartu, Department of Psychology Rossier, Jerome; University of Lausanne, Institute of Psychology Zecca, Gregory; University of Lausanne, Institute of Psychology Ah-Khion, Jennifer; University of Mauritius, Department of Social Studies Amoussou-Yéyé, Denis; University of Abomey-Calavi, Department of Psychology and Educational Sciences Bäckström, Martin; Lund University, Psychology Barkauskiene, Rasa; Vilnius University, Department of Clinical and Organizational Psychology Barry, Oumar; Cheikh Anta Diop University of Dakar, Faculty of Arts and Humanitary Bhowon, Uma; University of Mauritius, Department of Social Studies Björklund, Fredrik; Lund University, Department of Psychology Bochaver, Aleksandra; Moscow State University of Psychology and Education, Laboratory of Health Psychology Bochaver, Konstantin; Moscow State University of Psychology and Education, Department of Social Psychology de Bruin, Deon; University of Johannesburg, Department of Industrial Psychology and People Management Cabrera, Helena; University of Santo Tomas, College of Commerce and Business Administration Chen, Sylvia; The Hong Kong Polytechnic University, Applied Social Sciences Church, A.; Washington State University, Department of Educational Leadership and Counseling Psychology Cissé, Daouda; University of Bamako, Faculté des Lettres Dahourou, Donatien; University of Ouagadougou, Laboratoire de Psychologie Expérimentale Feng, Xiaohang; ChangChun Normal University, Department of

	Psychology Guan, Yanjun; Renmin University of China, School of Labor and Human Resources Hwang, Hyi-Sung; San Francisco State University, Department of Psychology Idris, Fazilah; Universiti Kebangsaan, Center for General Studies Katigbak, Marcia; Washington State University, Department of Educational Leadership and Counseling Psychology Kuppens, Peter; University of Leuven, Department of Psychology Kwiatkowska, Anna; Polish Academy of Sciences, Institute of Psychology Laurinavicius, Alfredas; Mykolas Romeris University, Department of Psychology Mastor, Khairul; Universiti Kebangsaan, Center for General Studies Matsumoto, David; San Francisco State University, Department of Psychology Riemann, Rainer; Bielefeld University, Department of Psychology Schug, Joanna; Hokkaido University, Department of Behavioral Science Simpson, Brian; San Francisco State University, Department of Psychology Tseung, Caroline; University of Mauritius, Department of Social Studies
Manuscript Keywords:	Cross-cultural research, Perception, Statistical methods

Running head: Comparability of Self-Reports across Cultures

Comparability of Self-Reported Conscientiousness across 21 Countries

René Mõttus[1,2], Jüri Allik[1,3], Anu Realo[1], Helle Pullmann[1], Jérôme Rossier[4], Gregory Zecca[5], Jennifer Ah-Kion[6], 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], Deon de Bruin[11], Helena F. Cabrera[12], Sylvia Xiaohua Chen[13], A. Timothy Church[14], Dougoumalé Cissé[15], Daouda Donatien Dahourou[16], Xiaohang Feng[17], Yanjun Guan[18], Hyi-Sung Hwang[19], Fazilah Idris[20], Marcia S. Katigbak[14], Peter Kuppens[21], Anna Kwiatkowska[22], Alfredas Laurinavicius[23], Khairul Anwar Mastor[20], David Matsumoto[18], Rainer Riemann[24], Joanna Schug[25], Brian Simpson[19], Caroline Ng Tseung[5]

[1] University of Tartu, Estonia
[2] University of Edinburgh, Centre for Cognitive Ageing and Cognitive Epidemiology
[3] Estonian Academy of Sciences, Estonia
[4] University of Lausanne, Switzerland
[5] University of Mauritius, Mauritius
[6] University of Cotonou, Benin
[7] Lund University, Sweden
[8] Vilnius University, Lithuania
[9] University of Cheikh Anta Diop, Senegal
[10] Moscow State University of Psychology and Education, Russia
[11] University of Johannesburg, South-Africa
[12] University of Santo Tomas, Philippines
[13] Hong Kong Polytechnic University, Hong Kong, People's Republic of China
[14] Washington State University, United States
[15] University of Bamako, Mali
[16] University of Ouagadougou, Burkina Faso
[17] ChangChun Normal University, People's Republic of China
[18] Renming University, People's Republic of China
[19] San Francisco State University, United States

[20] Universiti Kebangsaan Malaysia, Malaysia
[21] University of Melbourne, Australia, University of Leuven, Belgium
[22] Polish Academy of Sciences, Poland
[23] Mykolas Romeris University, Lithuania
[24] University of Bielefeld, Germany
[25] Hokkaido University, Japan

Corresponding author:
René Mõttus
University of Tartu
Tiigi 78, 50410 Tartu
Estonia
Phone: +372 7375903
Fax: +372 7376 152
E-mail: rene.mottus@ut.ee

Abstract

In cross-national studies, mean levels of self-reported phenomena are often not congruent with more objective criteria. One prominent explanation for such findings is that people make self-report judgements in relation to culture-specific standards (often called the reference group effect), thereby undermining the cross-cultural comparability of the judgements. We employed a simple method called *anchoring vignettes* in order to test whether people from 21 different countries have varying standards for Conscientiousness, a Big Five personality trait that has repeatedly shown unexpected nation-level relationships with external criteria. Participants rated their own Conscientiousness and that of 30 hypothetical persons portrayed in short vignettes. The latter type of ratings was expected to reveal individual differences in standards of Conscientiousness. The vignettes were rated relatively similarly in all countries, suggesting no substantial culture-related differences in standards for Conscientiousness. Controlling for the small differences in standards did not substantially change the rankings of countries on mean self-ratings or the predictive validities of these rankings for objective criteria. These findings are not consistent with mean self-rated Conscientiousness scores being influenced by culture-specific standards. The technique of anchoring vignettes can be used in various types of studies to assess the potentially confounding effects of reference levels.

KEYWORDS: anchoring vignettes; references group effect; DIF; cross-cultural; aggregate personality scores

Comparability of Self-Reported Conscientiousness across 21 Countries

Verbal self-reports are the most frequently used and sometimes the only available method in the social and behavioural sciences, health surveys, and other disciplines to collect information about how people feel or think or how they are expected to behave in certain situations. Self-reports are often employed to compare individuals within particular cultural settings, but they are also used for cross-national comparisons. For example, they form the basis of many types of international and regional rankings. At the same time, it is widely recognized that self-reports are prone to various errors and biases, such as self-enhancement and acquiescent responding (Church, 2009; Smith, 2004), which can influence comparisons both within and between different cultural settings. In the present study, we focus on a widely acknowledged problem related to comparing self-reports across cultures, *the reference group effect* (RGE; Heine, Lehman, Peng, & Greenholtz, 2002), and demonstrate means for both identifying and mitigating the problem.

It has been observed in psychology as well as several other disciplines that rankings of nations based on self-reports are not always congruent with relevant objective criteria. For example, when asked "How much say do you have in getting the government to address issues that interest you?" Chinese respondents tend to give higher ratings than Mexicans (King, Murray, Salomon, & Tandon, 2004), in spite of the fact that Mexico is ranked 81 positions higher than China on *The Economist* Democracy Index (The Economist, 2010). Likewise, in the field of health surveys, Sen (2002) showed that the prevalence of self-reported acute medical conditions is higher in regions where people, in fact, live longer and have better health. In psychology, it has been demonstrated that cross-cultural differences

in the individualism-collectivism dimension based on self-reports do not match with expert-rated differences in these cultures (Heine et al., 2002; but see also Takano & Sogon, 2008).

Another relevant example in comparative cultural research is related to personality traits. At the cross-national level, self-ratings of personality traits generally demonstrate a replicable pattern of geographic distribution (Allik & McCrae, 2004; Schmitt, Allik, McCrae, & Benet-Martinez, 2007), but some country rankings look strikingly counterintuitive. In particular, it is puzzling that inhabitants of countries with modest economic wealth, short life expectancy, low work-speed, and a high level of corruption perceive themselves as being more conscientious—determined, strong-willed, organized, dutiful, and deliberate—compared to people in more developed countries (Heine, Buchtel, & Norenzayan, 2008; Mõttus, Allik, & Realo, 2010; Oishi & Roth, 2009). Within cultures, at the level of individuals, the relationships are more in line with intuition: conscientious people tend to live healthier and longer lives (Bogg & Roberts, 2004; Kern & Friedman, 2008), have more successful careers (Judge, Higgins, Thoresen, & Barrick, 1999), and are less inclined to engage in antisocial behaviour (Miller & Lynam, 2001).

The lack of convergence between findings at the culture level and the individual level may be readily explainable, however, and it is often possible to find a sound theoretical explanation for this sort of discrepancy. A classic example of this is Robinson (1950). At the state-level in the US, a strong negative correlation (r = −0.53) was observed between the illiteracy rate and the proportion of the population born outside the US. Conversely, at the level individuals, the correlation was weakly positive (r = 0.12), showing that immigrants tended to have a higher illiteracy rate than native-born people. An obvious

explanation for this apparent paradox, also known as the "ecological fallacy," is that immigrants, who formed a small fraction of the total population, tended to settle in the states where the permanent population was more educated and perhaps more tolerant towards immigrants. With respect to Conscientiousness, however, we do not have a good explanation, as yet, why this should be higher in countries with less economic resources, lower life-expectancy, and higher corruption. Therefore, there are no reasons to rule out *a priori* the possibility that national mean scores of Conscientiousness reflect something else than the typical values of the trait within nations—that is, they might be biased.

Social comparison processes may provide one key explanation for the possibly paradoxical relationships between self-ratings and objective culture-level criteria. According to Leon Festinger's classical idea, people estimate their attitudes or dispositions relative to social standards (Festinger, 1954). For example, when people are asked how punctual they are, they are likely to formulate their answers in relation to generally accepted societal standards of punctuality. The problem is that these standards may systematically differ across cultures. Frequent travellers have probably noticed that "being on time" may mean arriving within a few minutes of schedule in one country, whereas a much greater leeway may be the norm in another country—an observation backed by recent scientific data (White, Valk, & Dialmy, 2011). Therefore, when people in various countries compare themselves to what is considered normative in their cultural context, their self-ratings can—partially or even mainly—differ because of varying reference standards (Heine et al., 2008). In other words, people in different cultures may translate identical trait-related information into completely different self-reports. Largely, this is similar to

what is often called differential item functioning (DIF). In the psychological literature, one such social comparison process has become known as the RGE (Heine et al., 2002).

The RGE in Cross-National Comparative Studies

The existence of the RGE has typically been demonstrated by varying the instructions given to respondents who fill out self-report measures and showing that these alterations result in different scores (e.g. Credé, Bashshur, & Niehorster, 2010; Heine et al., 2002; Oishi, Hahn, Schimmack, Radhakrishan, Dzokoto, & Ahadi, 2005). For instance, in a widely cited study by Heine and colleagues (2002), Canadians with Japanese cultural experience and Japanese with Canadian cultural experience were asked to complete an independence/interdependence scale with three different types of instructions: the first instruction did not emphasize any reference group, the second asked respondents to compare themselves to Japanese people, and the third one asked them to compare themselves to Canadians. The three different types of instructions resulted in different mean ratings whereas only the results from the opposite-culture reference group conditions (Canadians comparing themselves to Japanese people and vice versa) were consistent with the standard view about the differences between Canadian and Japanese cultures, according to which Canadians are more independent and less interdependent than Japanese.

However, the authors acknowledged that respondents may have based their perceptions of the specified reference groups on inaccurate stereotypes rather than on their actual knowledge about the members of the cultural groups, making the obtained group differences in independence and interdependence scores difficult to interpret (Heine et al., 2002). To mitigate this possibility, they asked people of European and Asian descent living

in Canada to complete the same measure without specifying any reference group. The researchers assumed that living in the same country would make the two groups of people rely naturally on the same reference group (although they admitted that this was probably not a fully correct assumption) and thereby provide comparable self-ratings. They again found support for the standard view—people with Asian ancestry were more interdependent and less independent. However, although these results have also been taken as a demonstration of the RGE, they in fact provide no *direct* evidence for it because the researchers did not actually test which standards the European and Asian Canadians had used in making their self-reports. It was merely an assumption (and, admittedly, not a completely correct one) that they had used the same standards: 'generic' Canadians. For instance, it was also possible that Asian Canadians had based their self-ratings on their (possibly inaccurate) stereotype of dominant European Canadians, again confounding the observed cultural differences.

Of course, these findings are likely to imply the existence of the RGE, which can confound cross-cultural comparisons of self-reports. However, it is evident that study designs based on manipulating instructions by explicitly specifying reference groups or employing multiple ethnic groups living in the same country inherently suffer from various significant limitations. The first limitation is precisely the one illustrated in the previous paragraph—uncertainty regarding the nature of respondents' perceptions of the reference groups specified in the instructions. Do people's perceptions of, say, Japanese reflect true population mean levels of the trait in question or are these perceptions just stereotypes that may or may not be accurate (McCrae, Terracciano, Realo, & Allik, 2007)? There is no solid evidence that individuals possess abilities to assess accurately how an average member of

the reference group thinks, feels, or behaves. As a result, when manipulating instructions by specifying different reference groups produces different results, this is neither direct nor incontrovertible evidence for the RGE. This may provide circumstantial but not definitive evidence for the RGE.

The second obvious limitation of these designs is that they are not readily usable in large-scale cross-cultural studies including numerous nations. Describing cultural variation more comprehensively than studies comparing only a few cultures, multinational research efforts are key contributors to cross-cultural personality psychology. Therefore, the multinational studies are precisely the area where addressing the potential confounding effects of the RGE is most important. The problem is that typical RGE study designs need people with multicultural experiences (Oishi et al., 2005). If people do not have enough firsthand experience or knowledge of the cultures in question, their perceptions of the specified reference groups will be based mostly on stereotypes. Obviously, however, people can have sufficient experience of only a limited number of cultures. Additionally, people with multicultural experiences are seldom representative members of their own cultures, further threatening the validity of the results.

In sum, the evidence reviewed above shows that there may be culture-related differences in the standards on which people base their self-report judgements of various traits and this may seriously confound cross-cultural comparability of self-reports. However, there is an urgent need for methods that would allow researchers to address the RGE problem without relying on potentially inaccurate stereotypes or involving exclusively

people with multicultural experience, and that would be readily employable in large-scale cross-cultural research.

A Potential Remedy for the RGE—Anchoring Vignettes

There is a discrepancy in current cross-cultural personality research that needs to be emphasized. The existence of the RGE is widely acknowledged and has almost become a truism (Church, 2009, 2010; Matsumoto & Yoo, 2006). Yet, when it comes to the currently influential large-scale cross-cultural personality studies that arguably define the field (e.g., De Fruyt, De Bolle, McCrae, Terracciano, & Costa, 2009; McCrae, Terracciano &78 Members of the Personality Profiles of Cultures Project, 2005; Schmitt et al., 2007), there has been little success or interest in addressing the problem. The reason for this disparity obviously lies in the fact that there have been no cost-effective methods for quantifying the RGE—potential differences in the standards on which people base their self-reports. Hence, the RGE has remained an abstract and impending threat that has not been adequately addressed. We believe, however, that a potential solution is available. In particular, a simple technique called anchoring vignettes (King et al., 2004)—originally developed outside of psychology—is applicable for the purpose of identifying differences in how people translate identical trait-related information into subjective self-reports—the very core of the RGE problem. Furthermore, the technique provides a means for *correcting* self-reports for potentially differing reference standards.

An Overview of the Anchoring Vignettes Technique

The fundamental idea of the anchoring vignettes technique is extremely simple (Hopkins & King, 2010; King et al., 2004; King & Wand, 2007). In a typical cross-cultural

study, respondents rate a phenomenon that is expected to vary across people and cultures (e.g., personality, values, or attitudes). Therefore, it is difficult, if not impossible, to identify whether their ratings differ because of true variance in the phenomenon or simply because people in different cultures endorse the questionnaire items in a different manner (e.g., due to the RGE). The anchoring vignettes technique allows researchers to estimate the latter type of variance by asking all respondents to rate something identical. The assumption is that if everyone rates the same target—or a set of targets—the only source of variance in their ratings can be biases or measurement error. Having quantified the (non-random) unwanted variance in the ratings (e.g., difference in the degree to which people endorse all items tapping a phenomenon, irrespective of the target of their ratings), the ratings can be corrected accordingly, resulting in bias-free ratings.

Obviously, it is important for the always-identical targets to be relevant to the phenomena being investigated. To achieve this, it is suggested that researchers create and administer to respondents, along with self-report scales, brief descriptions of hypothetical persons—the anchoring vignettes—that display various levels of the same characteristic being measured (e.g., political efficacy, perceived health, or Conscientiousness). If members of different groups have different standards for the trait being measured, there will be systematic group differences in the ratings of these vignettes. Assuming that vignette-ratings and self-reports are based on similar standards (e.g., the cultural norms for the trait), this would indicate that self-reports obtained from different groups are not directly comparable—exactly as the RGE predicts.

Importantly, the technique of anchoring vignettes is not limited to identifying differences in standards—it also provides a means for "fixing" the problem. If the vignettes are rated using the same scale people use to give their self-ratings (or any other type of rating that varies across people and cultures, such as peer-ratings), taking the difference between the two will result in standard-free self-ratings. In particular, self-ratings can be recoded to reflect their relative position among the hypothetical persons depicted in the vignettes (King & Wand, 2007), so that people's positions on the trait can vary from being lower than that of the lowest scoring hypothetical person to being higher than the highest scoring hypothetical person. Essentially, this recoding procedure means anchoring self-ratings to a "benchmark" common to all respondents.

Last but not least, it should be noted that the idea of anchoring self-ratings to specific hypothetical circumstances that are similar to all respondents is not new (e.g., Peng, Nisbett, & Wong, 1997). However, what is specific to and a strength of the anchoring vignettes method is the possibility of straightforward quantification of the RGE (in addition to the possibility of correcting self-ratings for its effect) by asking all respondents to rate the same targets.

RGE and More Traditional Approaches to Measurement (In)variance

Cross-cultural researchers have been concerned whether their multiple-item instruments work in the same way across cultures for quite some time already and tested for what is typically referred to as measurement invariance (MI). Undeniably, establishing MI is an inevitable precondition for scores of multiple-item instruments—presumably tapping a latent trait—to be comparable across groups (Meredith, 1993). However, it must be noted

that, compared to the RGE, MI is a conceptually different issue in cross-group comparisons. Specifically, MI addresses the degree to which indicators (items) contribute to a latent trait in the same way in different groups (with the same loadings, intercepts, and residual variances). In the core, establishing MI is a factor analytical procedure which taps the *relative* endorsement levels of items. The RGE, on the other hand, addresses whether people translate the same levels of a trait into the same *absolute* rating scores. That is, the RGE is basically a property of single items but, importantly, it can generalize across many items, thereby substantially affecting mean scores of multiple-item instruments. In particular, it is a realistic possibility that the RGE applies to all of the items of a single trait in the same way and to the same degree; for instance, due to some cultures having more lenient standards for every aspect of Conscientiousness than others. If this is true, MI procedures are not able to detect RGE, as it does not affect the relative contribution of items to the measurement of the latent trait. It only confounds mean levels of the traits.

Thus, the RGE is essentially a subtype of DIF. It may be argued that various procedures to detect DIF already exist (e.g., those based on item response theory). However, it is important to realize that there is a fundamental difference between the vignette-based procedure of detecting biases in ratings and the traditional DIF procedures. Namely, the vignettes provide an *external* "benchmark" (i.e., something other than the presumably substantive variation between individuals on the latent trait) against which to compare items to detect biases, whereas the other procedures rely on plotting single item scores against latent trait scores derived from basically the same type of information (e.g., using items from the same or similar scales). The problem is that when there is something systematically wrong with the type of information that we can obtain with this type of

ratings—such as an RGE present for all manifestations of the trait—the scores on the latent trait are affected in the same way than single items scores and the standard DIF detection procedures (similarly to MI procedures) do not identify the bias. Arguably, the inherent independence between the variance of the items in which DIF is tested and the (in)variance of the "benchmark" against which DIF is tested gives the vignette-based procedure an advantage over traditional DIF-detection procedures.

Aims of the Study

The anchoring vignettes technique is increasingly popular in comparative health (e.g., D'Uva, Van Doorslaer, Lindeboom, & O'Donnell, 2008), political (e.g. King et al., 2004), and economic research (e.g., Kristensen & Johansson, 2008) but is seldom employed in many other fields, including cross-cultural (or) personality psychology. However, we believe that it could be used to shed light on the afore-described puzzling problem of cross-cultural differences in personality ratings. Accordingly, the current study sets out to investigate the effect of potentially differing subjective standards on national rankings of different facets of self-reported Conscientiousness, the personality trait that has repeatedly shown unexpected national-level relationships with supposedly relevant objective criteria such as economic output or life-expectancy (Heine et al., 2008; Mõttus et al., 2010; Oishi & Roth, 2009). More specifically, using data from 21 different countries, we first studied the extent to which participants' country membership influenced their ratings on 30 anchoring vignettes that depicted hypothetical people with various levels of Conscientiousness. This initial analysis could potentially demonstrate the presence of an RGE-type phenomenon. Next, we investigated whether the differences in reference standards, as revealed by the anchoring vignettes, were likely to affect cultural rankings based on self-reports and

whether recoding participants' responses in relation to their ratings of hypothetical people had any actual effect on cultural rankings. Finally, we tested whether the corrected rankings of cultures predicted objective country-level criteria differently than the uncorrected rankings. In order to keep the RGE apart from other issues related to the comparability of ratings, such as absence of MI of latent traits (which were not the focus of this study), we carried all analyses out at the level of single items.

Method

Participants

Overall, 2,965 people from 21 countries took part in the study. The Peoples' Republic of China was represented with three independent samples—from Beijing, Changchun, and Hong Kong—but due to its high degree of autonomy and differing recent history, Hong Kong was treated as a separate country. The other two Chinese samples were tested with independently translated testing materials, leading us to treat them separately in all statistical analyses as well. The 22 samples consisted exclusively of university students in order to keep the demographic profiles of the samples as similar as possible. In the pooled sample, the mean age of participants was 22.17 years (SD = 5.27 years; range = 16 to 66 years) and 62.56% of the participants were woman. The demographic characteristics of the local samples are given in Table 1.

Table 1 about here

Testing Materials and Procedure

There is evidence that only some of the facets of Conscientiousness have counterintuitive cross-cultural rankings (Mõttus et al., 2010). For this reason, and in order to increase the likelihood of discovering the effects of subjective standard differences, we separately examined the different facets of Conscientiousness. We followed one of the most comprehensive models of Conscientiousness, the Five-Factor model of personality (FFM; McCrae & John, 1992), which describes this trait by way of six facets: Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, and Deliberation.

For each of the facets of Conscientiousness, five short descriptions of hypothetical people (vignettes) displaying various levels of the traits were drafted (the vignettes are given in Appendix I). The five hypothetical persons were intended to display very different levels of the trait, from very low to very high. The vignettes were first written in English. For cultures that use a primary language other than English, the vignettes—as well as all other testing materials—were carefully translated into the local language (and the names of the hypothetical people were changed to reflect cultural circumstances better). For each translation, independent back-translations into English were carried out and reviewed by the authors of the study. Where necessary, modifications were made.

Ideally, all vignettes should have described as specific and concrete behaviours as possible. However, it quickly became clear that this goal was not fully achievable as specific behaviours may have vastly different psychological and social meanings in different cultures (we emphasize that the present study incorporated a variety of cultures from nearly all continents). With that in mind, the vignettes were designed with an aim to

balance being specific enough and being applicable in each and every culture used in the study. Some of the vignettes referred only to specific and contextualized behaviours or life-achievements, while others were more abstract and decontextualized. Such variety among vignettes allowed for later selection between them, as well as for patterns in the findings to emerge (e.g., more concrete vignettes pointing to possible RGE but more abstract vignettes not).

Each of the six Conscientiousness facets was measured using a bipolar rating scale with the negative side of the trait described on one end of the scale and the positive side on the other (Terracciano et al., 2005). For instance, for the Competence facet, participants had to rate, on a five-point scale, their position between the end-points of the trait defined as "capable, efficient, competent" and "inept, unprepared." First, all participants rated their own personality using the six facets of Conscientiousness. Second, all respondents rated all hypothetical people in the 30 vignettes using the same set of bipolar rating scales. Finally, respondents provided information about their demographic background including age and sex.

Controlling for the Effects of Age and Sex

There was some heterogeneity among samples in terms of mean age and the proportion of women (Table 1). At the same time, small but fairly universal age and gender differences have been observed in Conscientiousness (McCrae, Terracciano, & 78 Members of the Personality Profiles of Cultures Project, 2005), and it was also possible that age and sex may be related to standards applied in vignette-ratings. Therefore, to avoid the confounding effects of age and sex proportion differences between samples, we adjusted all ratings for

raters' age and sex. First, the linear effects of age on all ratings were calculated and, using the regression parameter, ratings were transformed so that they were as if they all had belonged to 20-year olds. As the next step, gender differences were removed from the age-adjusted ratings.

Choosing the Best Combinations of Vignettes

Before recoding the self-ratings, we examined the sets of vignettes written for each facet for their ability to produce the most informative recodings of respondents' self-ratings. Generally, the more vignettes researchers have for correcting a particular self-rating, the greater the number of categories that the self-ratings can be sorted into and, as a result, the higher the discriminatory power of the recoded self-ratings (King & Wand, 2007). However, a higher number of vignettes also brings about a higher likelihood that the vignettes will be rated inconsistently: some respondents may deviate from the expected ranking of vignettes by giving two vignettes an equal rating, or rate the vignettes in a way that contradicts the expected ranking altogether. In these cases, the recoding does not produce a single (scalar) value for the respondent's self-rating but rather a range of possible values (vectored value) (King & Wand, 2007). Such vectored values can be used in various statistical analyses. However, as they contain less exact information than scalar values, it is reasonable to reduce their prevalence in the first place. Therefore, when deciding on the optimal set of vignettes, there is a trade-off between the level of informativeness and the number of vectored values that results from employing any particular set of vignettes. In order to quantify the level of informativeness of any set of vignettes, King and Wand (2007) have developed a formal measure called *entropy*. The set with the lowest entropy is the one that sorts every respondent into the minimal number of categories, whereas the

highest entropy characterizes the set of vignettes that sorts people equally into all categories.

When choosing the optimal set of vignettes, we balanced entropy with the minimum number of recoded self-ratings having vectored values. For calculating entropy, software developed by Wand, King, and Lau (in press) was used. These analyses were done on ratings unadjusted for age and sex differences because sex and age were included as co-variates in the entropy models. Generally, each additional vignette added increasingly less information. Having five vignettes instead of four added only little entropy, the same being generally true when four vignettes were used instead of three. The reason for some vignettes being relatively less informative than others was that they reflected trait levels that were either too low or too high and therefore only a few people could have been recoded around them (e.g., having a value that is lower than that of the lowest scoring hypothetical person). At the same time, having three vignettes instead of two increased entropy considerably. Therefore, we chose sets of three vignettes for all facets, balancing high entropy with as low number of vectored values as possible (retained vignettes are indicated in Appendix I). After recoding the self-ratings using the chosen sets of vignettes, 84, 90, 65, 83, 92, and 68 percent of the recoded self-ratings had scalar values for Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, and Deliberation, respectively.

Censored Ordered Probit Model

Thus, although a majority of the recoded values was scalar, we also had to deal with vectored values. Fortunately, the censored ordered probit model (COP), a generalisation of the standard ordered probit model (SOP) developed by King and Wand (2007), is able to

use vectored values in addition to scalar values as dependent variables. In scalar values, the COP acts exactly as the SOP, whereas in vectored values it collapses all the response categories spanned by the vector into a single category (Hopkins & King, 2010). Importantly, the regression coefficients from the COP are interpretable exactly in the same manner as those from the SOP (or any other linear regression model). SOP and COP regressions were carried out using an R-package developed by Wand, King, and Lau (in press). SOP and COP analyses were carried out on unadjusted ratings, as sex and age were used as co-variates in the models.

Results

Sample-level means and standard deviations on the six facets of self-rated Conscientiousness are given in Table 2. Full data are available on request from the first author.

Table 2 about here

Individual Differences in Conscientiousness Were Perceived Similarly across Countries

We first addressed possible cross-sample differences in how people perceived the differences between the hypothetical people. In addition to possible differential endorsement levels of personality ratings (e.g., due to the RGE), an important assumption for personality ratings to be comparable across groups is that *individual differences* on the traits are perceived and rated similarly. If the same people are ranked differently in different groups, this would also imply major problems for the comparability of the ratings. However, this appeared not to be the case. Differences in the levels of Conscientiousness between the hypothetical individuals were rated very similarly across samples. Sample-

level profiles consisting of the mean ratings of the 30 vignettes (22 profiles, one for each sample) were highly similar, with Spearman rank-order correlations between them ranging from 0.83 (between Benin and Japan) to 0.98 (between Australia and the USA, Germany and Sweden, and Switzerland and the USA), with a median of 0.93. This suggests that in relative sense personality ratings were fairly universal—relatively higher levels of Conscientiousness tended to be rated higher everywhere, and relatively lower levels of the trait tended to be universally rated lower.

Sample-Related Variance in Self- and Vignette-Ratings

Consistency in the rankings of the anchoring vignettes does not preclude substantial differences in the mean levels of the ratings: although individual *differences* were perceived similarly across cultures, they could have been translated into ratings with different endorsement levels, which is the very prediction of the RGE. To investigate this possibility, we examined the degree to which cultural background affected the overall variability in the ratings of the anchoring vignettes. Certainly not everyone rated the anchoring vignettes identically (all vignette-ratings had variances far above zero) but the crucial question was how much of the variability could be ascribed to the respondents' sample of origin. A one-way ANOVA revealed that the eta-squares ranged from 0.02 to 0.10 across the 30 anchors, with a median of 0.04. That is, on average, 4% of the overall variability in the anchor ratings could be ascribed to the differences in sample means. However, in order to more meaningfully interpret the degree of culture-related variance in the vignette-ratings, we compared it to the corresponding variance in self-ratings.

In particular, if it is true that people rate themselves wholly in relation to culture-specific standards, then mean self-ratings should vary across cultures only as much as the standards vary. Translating this into the present context, if the RGE had been able to reverse the rankings of cultures on self-reported Conscientiousness, we would have expected the differences in sample means of vignette-ratings to be at least as large as the differences in mean self-ratings. However, this was not the case—self-ratings in fact varied more across samples than vignette-ratings. For the six facets of Conscientiousness, eta-squares quantifying sample-related variance in self-ratings ranged from 0.07 to 0.13, with a median of 0.09. Thus, the sample-related variability in self-ratings was, on average, about twice as large as the variability in vignettes-ratings.

Sample-Level Associations between Vignette- and Self-Ratings

As an interim summary, respondents from different cultures ranked personality differences between people in much the same way and rated themselves to be more different than they rated the always-identical hypothetical persons described in the vignettes. These findings are necessary—but not sufficient—preconditions for self-reports to be comparable across cultures without the confounding effect of the RGE. The next important question, however, was whether the cross-sample differences in the vignette-ratings—despite being small—were in the same direction as the cross-cultural differences in self-ratings. If the reference standards underlying the RGE indeed differed across samples and could, in principle, alter rankings on self-rated Conscientiousness, they should have influenced self- and vignette-ratings in the same way. That is, due to harsh standards in some cultures, people should have rated themselves low and they should have also rated everyone else low, including the hypothetical persons depicted in the vignettes; the reverse

should also be true—in some cultures, lenient standards for the trait should have lifted all ratings, regardless of the target.

However, this was not the case. Table 3 gives the rank-order correlations between mean self-ratings and the mean ratings given to the vignettes of the respective facets. There was no systematic trend for mean self- and vignette-ratings to be in the same direction. Only 6 of the 30 correlations were statistically significant at any traditional alpha level (i.e., $p < 0.05$ or lower), with exactly half of them being negative. We take this as one of the indications that the rankings of samples on self-rated Conscientiousness were probably not substantially or systematically affected by differences in the subjective standards people had based their ratings on.

Table 3 about here

The Effect of Correcting for the RGE on the Rankings of Samples

We further attempted to quantify the possible effect of differences in reference standards on self-reports by making full use of the anchoring vignettes technique and directly comparing the rankings of samples on uncorrected self-ratings to the rankings on self-ratings that were corrected using the vignettes. Firstly, we ran six SOP regressions, predicting raw self-ratings on each of the six facets of Conscientiousness by respondents' sample-membership, age, and sex. Resulting regression coefficients could effectively be used to rank samples on the basis of *uncorrected* scores on Conscientiousness facets. Secondly, we ran six COP regressions on the recoded self-ratings of the facets, again using

sample-membership, age, and sex as predictors. Now, the resulting regression coefficients could be used to rank samples on the basis of *corrected* self-ratings.

Having the two rankings (Table 4), we could formally investigate the degree to which they overlapped. Although not identical, the uncorrected and corrected sample rankings appeared to be highly similar, with the rank-order correlations between them ranging from 0.78 (Self-Discipline) to 0.93 (Achievement Striving) across the six facets of Conscientiousness (the median correlation was 0.86). The biggest changes in rankings were for Estonia, which raised 10 positions on Dutifulness after correction, and Hong Kong, which declined 10 positions on Self-Discipline. In most cases, however, samples moved less in the rankings, shifting approximately two positions up or down, on average. The relatively modest effect of correcting self-ratings is not consistent with the results of cross-cultural comparisons on Conscientiousness being substantially influenced by differences in the ways in which people translate trait-related information into response categories of rating scales.

Table 4 about here

The Effect of Correcting for the RGE on Predictive Validity

Finally, although the effect of correcting self-ratings for differences in standards appeared to be fairly small, we examined whether it influenced the predictive validity of mean personality trait scores in any direction. In particular, it has to be borne in mind that correlations are non-transitive. For example, if uncorrected rankings on self-ratings are correlated with a criteria with a value of 0.50 (which is a rather high expectation in this

context; see Table 3 in Mõttus et al., 2010), then, unless the correlations between corrected and uncorrected rankings are greater than 0.86 (the observed median in this study), the correlations of corrected rankings with the criteria do not necessarily have to be higher than zero.

Since country-level mean Conscientiousness scores have—for many people, unexpectedly—shown negative relationships with longevity and national wealth (Heine et al., 2008; Mõttus et al., 2010; Oishi & Roth, 2009), we compared the degree to which the uncorrected and corrected rankings of samples on the facets of Conscientiousness (Table 4) predicted countries' life expectancies and Gross Domestic Product (GDP) per capita. Consistent with the previous studies, uncorrected country rankings on Conscientiousness facets related negatively to life expectancy and GDP (Figure 1). After correcting the self-ratings, the relationships remained negative, although the correlations were to some extent weaker for some facets. These results showed that the counterintuitive relationships between country-level mean Conscientiousness scores and their supposedly relevant objective criteria probably did not result from culture-specific standards that people had referred to when giving personality ratings.

Figure 1 about here

Discussion

In several published studies, the technique of anchoring vignettes has successfully identified the RGE on cross-cultural rankings of self-reported phenomena such as political beliefs and work satisfaction (e.g., King et al., 2004; Kristensen & Johansson, 2008).

However, applying the technique to Conscientiousness—the personality trait that has shown puzzling cross-cultural rankings in previous studies and could therefore possibly suffer from an RGE-type measurement confounding (Heine et al., 2008; Mõttus et al., 2010; Oishi & Roth, 2009)—we were not able to reveal any substantial effect of culture-specific standards on the ranking of countries or the predictive validity of these rankings. This was separately tested for six facets of Conscientiousness by using 30 independent vignettes and the results, indicating only a minor effect of culture-specific standards, were fairly robust. Although the current implementation of the anchoring vignettes technique may possibly have some important limitations, as will be discussed below, we tend to believe that mean self-rated Conscientiousness scores do not suffer from culture-specific standards for the trait. We now turn a discussion of the implications of this conclusion.

What Might Be Going on with the Country-Level Mean Scores of Conscientiousness?

The conclusion that the RGE may have only a limited effect on self-rated Conscientiousness scores leaves us with two broad groups of explanations with regard to national rankings of the trait. First, despite the modest effect of the RGE, as suggested by the present findings, the national rankings may still be biased. That is, there may be factors other than the RGE that distort self-reports in cross-national comparisons and make the rankings counterintuitive. One of the factors may be differential self-enhancement, suggesting that, although people may refer to more or less universal standards when judging the various aspects of Conscientiousness, their motivation to present themselves (as opposed to other people, including the hypothetical persons described in the vignettes) in a favourable manner (i.e., high on Conscientiousness) may differ across cultural settings. Indeed, there is some evidence that East-Asians tend to engage in self-enhancement

differently than Westerners (Heine, Kitayama, & Hamamura, 2007). On the other hand, a recent large-scale study found that the degree to which mean self-ratings on the NEO PI-R facets differ from mean observer-ratings on the same traits is fairly similar across a wide range of cultures (Allik et al., 2010). These findings suggest that the ratio of self-enhancement to other-enhancement on personality traits is relatively universal, making an enhancement-based explanation for the national rankings of personality traits less likely.

Another possible bias in nation-level personality scores may be related to selective sampling. In particular, most of the nation-level average self-reported personality scores are based on student samples (McCrae, 2002; Schmitt et al., 2007). While it is obvious that students are not likely to comprise perfectly representative samples of general populations (Henrich, Heine, & Norenzayan, 2010), their cross-national comparability may be further complicated by the possibility that in different countries students differ from the general population in different ways. For instance, in some countries it is easier to be admitted to university (e.g., free admission to everyone at the beginning, followed by a subsequent dropout of less successful students) than in other countries (e.g., strict admission requirements), which may automatically introduce selection bias. Due to these differences, it is possible that certain personality traits—high Conscientiousness possibly being one of them—are differentially advantageous in terms of being admitted to university, leading to cross-national differences in the proportion of highly conscientious people in universities. Some evidence for this explanation comes from the finding that national mean scores of observer-rated Conscientiousness which described more heterogeneous populations than students (McCrae et al., 2005) have shown slightly less counterintuitive correlations with potential objective criteria of the trait (Heine et al., 2008; Mõttus et al., 2010). However, it

is important to realize that if selective sampling is indeed the "problem" related to national mean scores of personality traits, this would in fact be good news for cross-cultural personality psychology, as recruiting more representative samples is arguably a far simpler task than battling with the obscure inherent biases in self-reports such as the RGE.

The second broad explanation for the national rankings on Conscientiousness is that the rankings more or less accurately reflect real differences between nations but researchers' intuitions about Conscientiousness or its relationships to objective criterion variables have been inaccurate (Mõttus et al., 2010). Given our currently limited understanding of the culture-personality interface, we have to acknowledge the possibility that even the seemingly most reasonable predictions about the relationships between self-reported personality scores and other country-level variables may ultimately prove to be untenable. For instance, the studies described above expected nation-level mean Conscientiousness scores to be positively correlated with nations' economic output, operationalized as GDP per capita. This expectation has probably been based on individual-level findings which tend to show that high Conscientiousness is related to just about every socially valued outcome, including being economically successful. However, proposing similar links at the level of cultures requires rigorous theoretical elaboration before they can be taken as an *a priori* correct assumptions (i.e., before a personality test's ability to reproduce these associations is viewed as the validity criterion of the test).

To illustrate the complexity of the associations between the average Conscientiousness of people and the relative amount of circulating money in a society (the GDP), we can imagine several radically different ways to think about the relationship (for a *prima facie*

illustration, see Hofstede & McCrae, 2004). First, we may assume that typical personality trait levels in a society cause the societal outcome. This is a perfectly plausible supposition, but it is important to realize that there are probably millions of reasons why societies differ with respect to the amount of money circulating in them, and the personality trait levels of their members constitute only one of the many, if at all. It seems highly likely that the currently available cross-cultural studies have been underpowered to reliably detect these presumably weak associations in the first place. Conversely, we may assume that the amount of wealth determines people's levels of Conscientiousness, with greater opportunities to earn and spend making people less reliable, disciplined and deliberate (Hofstede & McCrae, 2004; Hofstede's interpretation, p. 74). This is also a viable possibility but, again, individual and cultural differences in personality traits are likely to be influenced by a myriad of reasons, societal differences in economic output possibly being only one of them. Finally, we may assume that there are reciprocal effects between mean personality trait levels and societal indicators. However, predicting the nature of such relationships would presumably be an even more complicated endeavour than unpacking any unidirectional associations.

Limitations and Future Considerations

We note that the study has a potential limitation that may have influenced its findings in important ways. Namely, the purpose of including a wide array of cultures in the study, to cover as much cultural variability as possible, did set some limits with respect to drafting the vignettes, as mentioned above. The content of the vignettes had to have reasonably universal meanings across the cultures and therefore the vignettes often could not describe highly specific and contextualized behaviours. It may therefore be argued that the vignettes

did not provide enough solid "anchors" for subjective standards as people may have perceived the *content* of vignettes differently (which is different to translating the same content into different ratings because of different subjective standards for the trait—the very phenomenon we were testing for). Had this been true, the vignette-ratings may have differed across cultures not only due to the RGE but also due to differently perceived content, meaning that the variance in the vignette-ratings may have largely reflected noise. This, nonetheless, was not likely, as we observed remarkable regularity in the ratings (e.g., highly similar rankings of the vignettes across cultures and similarity between uncorrected and corrected self-ratings). Alternatively, it may be argued the vignette-ratings were not expected to vary across cultures because the vignettes were too abstract and vague for culture-specific standards to apply to them. Indeed, the vignette-ratings did not show much culture-related variance.

We acknowledge the fact that several vignettes were rather abstract. However, this was not true for all 30 of the vignettes. There was notable variability among the vignettes in terms of specificity and the degree of contextualization. One example of a vignette that refers to a specific behaviour is the following: "Alex' work day is rarely shorter than 12 hours and he had his last holiday 5 years ago. At work he tries to get additional assignments in order to be distinguished. Alex dreams about becoming the manager of his current institution" (#C4.2 in Appendix I). Yet, neither this nor *most* of the other concrete vignettes showed culture-related differences in the same direction as self-ratings, something that could have signalled a possible effect of the RGE on self-ratings. A clear exception, however, was vignette #C3.1 (Appendix I), which was extremely specific in content and, at the same time, showed a positive correlation ($r = 0.47$, $p < 0.05$) with the respective self-

ratings across cultures (see Table 3). In principle, this leaves open the possibility that using more specific and contextualized vignettes may potentially have resulted in different findings. Therefore, acknowledging the possibility that the vignettes used in this study were not always ideal for the purpose of providing solid anchors for subjective ratings, we urge future studies to make an extra effort to design vignettes at different levels of specificity.

It is also worthwhile pointing out that the anchoring method did not allow us to directly address possible cross-cultural differences in the relevance of various manifestations of Conscientiousness. It may have been that the content of the vignettes—however specific—was not equally relevant in each and every cultural setting. On the other hand, there is a substantial amount of literature showing that the structural properties of personality inventories tend to be replicable in a wide range of cultures (De Fruyt et al., 2009; McCrae et al., 2005; Schmitt et al., 2007), suggesting that the content of basic personality traits, including Conscientiousness, tends to be more or less similar across cultures. This gives us some confidence in the belief that the content of the vignettes was similarly relevant across all of the cultural settings covered in the study. Another reason to believe that differential relevance of the content of the vignettes was not a major problem was the robustness of the vignette-ratings: they were ranked similarly and endorsed largely to the same degree in all countries studied and produced recoded self-ratings that were similar to uncorrected self-ratings. Had the meaning of the vignettes substantially varied across cultures, we would have probably seen much less regularity in the ratings.

Apart from the content of the vignettes, future studies are likely to benefit from varying the order in which vignette-ratings and self-ratings are requested from respondents.

In the present study, self-ratings were given prior to rating the vignettes. Considering the possibility that presenting people with the vignettes may have influenced their subsequent self-ratings (e.g., by providing explicit comparison standards), only the present approach allowed the testing of the effect of potentially differing reference standards on "intact" self-ratings (i.e., as they would normally be obtained in any other study). In other words, if people's self-ratings had been obtained after presenting them with vignettes, the self-ratings might have already been influenced in a systematic way and therefore any results based on them (including the effect of correcting for the RGE) would have had limited generalizablity. However, it is important to note that the possibility that the method of presentation of vignettes can influence self-ratings is not necessarily negative. On the contrary, if presenting people with vignettes is sufficient to render their subsequent self-ratings more comparable—as was indeed recently demonstrated by Hopkins and King (2010)—this would provide another method for improving the validity of self-ratings, including their cross-cultural comparability. To combine the merits of both approaches, in future studies researchers are encouraged to collect vignette-ratings and self-ratings in both orders (e.g., by assigning respondents randomly into two groups with different orders of presentation). This would allow for the testing of whether the order of presentation has a systematic effect on the validity of the self-ratings or not.

Conclusions

This study represents an important step towards being able to empirically identify and handle what is often considered a major problem for cross-group comparability of personality ratings—the RGE. More specifically, the results of this study are not consistent with mean self-rated Conscientiousness scores being substantially influenced by the RGE.

However, further research is certainly needed to clarify this issue as one study can never be sufficient for definitive conclusions. Furthermore, this study may have suffered from methodological limitations, such as the use of too abstract and decontextualized vignettes. Additionally, future studies will have to show whether other personality traits are also likely to be judged in absolute rather than in relative terms. It is possible, for example, that people have developed a more robust and unconditional way to assess their basic tendencies to feel, think, and behave than to assess the level of political freedom in their society or their work satisfaction (King et al., 2004; Kristensen & Johansson, 2008). In much of their daily lives, people are surrounded by personality-relevant information and they constantly have to act on the basis of this information, probably leading them to be highly trained in making personality judgments about themselves and others. In sum, if the present findings can be replicated and are also found to apply to other personality traits, then ruling out the existence of the widely suspected confounder of personality self-reports—the RGE—will represent an important step towards being finally able to interpret observed cross-national personality differences in a substantive manner.

One important outcome of the study is the demonstration of a relatively easy technique for mitigating the potential RGE problem. Although this study focused exclusively on one specific personality trait, the problem of the possible incomparability of self-reports and the ways of addressing this problem have implications for many research areas in psychology. As demonstrated by the results of this study, the simple and cost-effective method of anchoring vignettes (King et al., 2004) can be routinely used in any kind of cross-national or comparative research involving self-reports. Importantly, the method is also applicable to areas other than cross-cultural research. For instance, if there are reasons to hypothesize

age- or education-related differences in the ways people use rating scales, the technique of anchoring vignettes can be easily used to deal with such differences.

References

Allik, J., & McCrae, R. R. (2004). Toward a geography of personality traits - Patterns of profiles across 36 cultures. *Journal of Cross-Cultural Psychology*, *35*, 13-28.

Allik, J., Realo, A., Mõttus, R., Borkenau, P., Kuppens, P., & Hrebícková, M. (2010). How people see others is different from how people see themselves: A replicable pattern across cultures. *Journal of Personality and Social Psychology*, *99*, 870-882.

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin*, *130*, 887-919.

Church, A. T. (2010). Current perspectives in the study of personality across cultures. *Perspectives on Psychological Science*, *5*, 441 -449.

Church, A. T. (2009). Prospects for an integrated trait and cultural psychology. *European Journal of Personality*, *23*, 153-182.

Credé, M., Bashshur, M., & Niehorster, S. (2010). Reference group effects in the measurement of personality and attitudes. *Journal of Personality Assessment*, *92*, 390.

De Fruyt, F., De Bolle, M., McCrae, R. R., Terracciano, A., & Costa, P. T. (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment*, *16*, 301-311.

D'Uva, T. B., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, *17*, 351-375.

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., Kitayama, S., & Hamamura, T. (2007). Which studies test whether self-enhancement is pancultural? Reply to Sedikides, Gaertner, and Vevea, 2007. *Asian Journal of Social Psychology*, *10*, 198-200.

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.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*, 29.

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

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

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*, 621-652.

Kern, M. L., & Friedman, H. S. (2008). Do conscientious individuals live longer? A quantitative review. *Health Psychology*, *27*, 505–512.

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.

Matsumoto, D., & Yoo, S. H. (2006). Toward a new generation of cross-cultural research. *Perspectives on Psychological Science*, *1*, 234-250.

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

McCrae, R. R., Terracciano, A., Realo, A., & Allik, J. (2007). On the validity of culture-level personality and stereotype scores. *European Journal of Personality, 21*, 987-991.

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.

McCrae, R. R., Terracciano, A. & 79 Members of the Personality Profiles of Cultures Project. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology*, *89*, 407–425.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Miller, J. D., & Lynam, D. (2001). Structural models of personality and their relation to antisocial behavior: A meta-analytic review. *Criminology*, *39*, 765-798.

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.

Oishi, S., Hahn, J., Schimmack, U., Radhakrishan, P., Dzokoto, V., & Ahadi, S. (2005). The measurement of values across cultures: A pairwise comparison approach. *Journal of Research in Personality, 39,* 299-305.

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.

Peng, K., Nisbett, R. E., & Wong, N. Y. C. (1997). Validity problems comparing values across cultures and possible solutions. *Psychological Methods, 2,* 329-344.

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

Schmitt, D., 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.

Sen, A. (2002). Health: perception versus observation. *British Medical Journal*, *324*, 860-861.

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

Takano, Y., & Sogon, S. (2008). Are Japanese more collectivistic than Americans? *Journal of Cross-Cultural Psychology*, *39*, 237 -250.

Terracciano, A., Abdel-Khalek, A. M., Adam, N., Adamovova, L., Ahn, C., Ahn, H. N., et al. (2005). National character does not reflect mean personality trait levels in 49

cultures. *Science*, *310*, 96–100.

The Economist Intelligence Unit's Index of Democracy 2008. Retrieved on February 28, 2010, from http://graphics.eiu.com/PDF/DemocracyIndex202008.pdf.

Wand, J., King, G., & Lau, O. (in press). anchors: Software for anchoring vignette data. *Journal of Statistical Software*.

White, L. T., Valk, R., & Abdessamad, D. (2011). What is the meaning of "on time"? The sociocultural nature of punctuality. *Journal of Cross-Cultural Psychology, 42, 482-493.*

Acknowledgements

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. The authors are grateful to Steven Heine for his helpful comments on a draft of the manuscript.

Table 1. Demographic Characteristics of Samples

	Language	N	% female	Mean age	SD of age	Age range
Australia	English	463	76.24	22.11	6.11	18-55
Benin	French	107	41.12	24.77	5.99	19-55
Burkina Faso	French	96	35.42	25.67	4.26	19-41
China (Changchun)	Chinese	110	78.18	27.99	3.56	22-37
China (Beijing)	Chinese	150	47.33	18.67	0.96	16-22
Estonia	Estonian	110	72.73	21.15	5.36	18-66
Germany	German	70	88.57	22.99	5.34	19-49
Hong-Kong	Chinese	158	51.27	20.58	1.58	18-30
Japan	Japanese	107	59.81	20.63	2.72	19-41
Lithuania	Lithuanian	125	68.80	19.02	0.93	18-25
Malaysia	Malay	211	69.19	19.82	1.38	18-30
Mali	French	93	23.66	28.84	6.95	20-50
Mauritius	French	100	48.00	20.69	2.21	18-35
Philippines	Filipino	133	55.64	18.60	0.81	17-21
Poland	Polish	100	84.00	24.46	5.92	20-50
Russia	Russian	100	57.00	18.73	1.93	16-24
Senegal	French	115	42.61	27.58	6.39	18-50
South Africa	English	109	68.81	20.36	2.87	17-31
South-Korea	Korean	142	57.04	22.10	2.31	19-27
Sweden	Swedish	100	52.00	25.23	2.87	20-35
Switzerland	French	101	74.26	20.89	3.53	18-38
USA	English	165	79.39	23.12	7.82	18-58

NOTE: SD = Standard deviation.

Table 2. Age- and sex-adjusted means and standard deviations of self-ratings.

	Competence		Order		Dutifulness		Achievement Striving		Self-Discipline		Deliberation	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Australia	3.84	0.97	3.38	1.13	3.86	0.88	3.46	0.94	3.34	1.00	3.55	1.01
Benin	4.38	0.62	4.20	0.93	4.42	0.74	4.38	0.77	4.16	1.00	4.21	1.05
BurkinaFaso	4.02	0.76	3.92	1.15	4.41	0.74	4.34	0.71	4.01	1.05	4.24	0.78
China (Changchun)	3.98	0.79	3.75	0.96	4.29	0.78	3.78	0.85	3.78	0.92	3.78	1.00
China (Beijing)	3.86	0.82	4.05	0.87	4.37	0.79	3.66	0.92	4.18	0.84	3.83	1.02
Estonia	3.87	0.76	3.49	0.95	4.07	0.89	3.50	0.89	3.41	1.06	3.89	0.91
Germany	3.86	0.75	3.78	1.01	3.83	0.74	3.69	0.85	3.73	0.94	3.39	0.99
Hong-Kong	3.52	0.96	4.10	1.06	3.77	0.92	3.91	0.97	3.31	0.97	3.98	0.97
Japan	2.99	0.99	2.79	1.14	3.44	1.05	3.10	1.18	3.09	1.07	3.27	1.10
Lihtuania	3.72	0.82	3.65	0.91	3.92	0.84	3.23	0.87	3.46	0.82	3.40	1.10
Malaysia	3.71	0.76	3.99	0.90	4.16	0.85	4.10	0.81	3.64	0.92	3.82	0.91
Mali	3.96	0.75	3.73	0.99	4.25	0.84	4.28	0.75	4.04	0.95	4.07	0.81
Mauritius	3.82	0.95	3.54	1.04	3.93	0.99	3.66	0.94	3.81	0.90	3.63	1.10
Philippines	4.10	0.70	3.80	0.94	4.15	0.70	4.15	0.78	3.94	0.91	3.84	0.99
Poland	4.11	0.87	3.95	0.92	4.18	0.82	3.67	0.92	3.61	1.09	3.54	1.08
Russia	3.98	0.86	3.64	1.02	3.77	1.10	3.64	1.10	3.41	1.12	3.77	0.94
Senegal	4.10	0.71	3.80	1.05	4.41	0.77	4.32	0.75	4.14	0.90	3.92	1.12
South Africa	4.31	0.97	3.73	0.98	4.13	0.91	4.05	0.90	3.92	1.21	4.00	1.14
South-Korea	3.49	0.96	3.37	0.98	3.78	0.89	3.32	0.98	3.00	1.01	3.62	0.93
Sweden	3.99	0.79	3.67	0.96	3.92	0.84	3.67	0.68	3.94	0.79	3.25	1.01
Switzerland	3.88	0.70	3.38	0.97	4.14	0.67	3.51	0.84	3.67	1.03	3.45	1.07
USA	4.31	0.69	3.73	0.88	4.01	0.79	3.69	0.87	3.80	0.95	3.69	0.92

NOTE: M = Mean score; SD = standard deviation.

Table 3. Spearman rank-order correlations between sample-level mean self-ratings and mean vignette-ratings of the same facets of Conscientiousness.

	Competence	Orderliness	Dutifulness	Achievement Striving	Self-Discipline	Deliberation
Vignette 1	-0.03	0.08	**0.47**	-0.26	0.22	-0.19
Vignette 2	-0.27	-0.39	0.39	-0.18	-0.05	0.12
Vignette 3	**-0.44**	0.03	-0.36	0.40	**-0.45**	**0.59**
Vignette 4	0.26	0.15	**0.56**	-0.32	-0.16	0.06
Vignette 5	0.14	-0.02	-0.06	0.42	-0.41	**-0.58**

NOTE: Correlations significant at $p < 0.05$ are given in bold. Vignettes are in the same order as in Appendix I

Table 4. SOP and COP Regression Parameter Estimates and Standard Errors of the Estimates for the Six Facets of Conscientiousness.

	Competence				Order				Dutifulness			
	SOP		COP		SOP		COP		SOP		COP	
	Estim	St Err	Estim	St Err	Estim	St Err	Estim	St Err	Estim	St Err	Estim	St Err
Benin	1.13	0.21	0.59	0.12	1.62	0.21	0.66	0.12	1.42	0.22	0.89	0.13
Burkina Faso	0.35	0.22	0.03	0.12	1.10	0.22	0.54	0.12	1.40	0.23	0.74	0.14
China (Changchun)	0.33	0.21	0.38	0.12	0.65	0.20	0.30	0.12	1.18	0.22	0.67	0.13
China (Beijing)	-0.28	0.18	-0.36	0.11	1.09	0.17	0.45	0.10	1.06	0.18	0.60	0.11
Estonia	-0.13	0.19	0.12	0.11	0.11	0.19	-0.03	0.11	0.49	0.20	0.04	0.12
Germany	-0.14	0.24	0.10	0.14	0.68	0.24	0.55	0.14	-0.13	0.23	0.15	0.14
Hong Kong	-0.76	0.17	-0.55	0.10	1.43	0.18	0.90	0.10	-0.19	0.17	0.09	0.10
Japan	-1.80	0.20	-1.18	0.12	-0.97	0.20	-0.36	0.12	-0.79	0.20	-0.36	0.12
Lithuania	-0.55	0.19	-0.27	0.11	0.37	0.18	0.02	0.11	0.05	0.19	-0.07	0.11
Malaysia	-0.70	0.16	-0.41	0.09	0.87	0.16	0.38	0.09	0.48	0.16	0.26	0.10
Mali	0.28	0.22	-0.13	0.13	0.71	0.21	0.15	0.13	1.08	0.23	0.56	0.14
Mauritius	-0.22	0.21	0.06	0.12	0.24	0.20	0.24	0.12	0.25	0.21	0.25	0.13
Philippines	0.28	0.19	0.01	0.11	0.66	0.18	0.15	0.11	0.44	0.18	0.16	0.11
Poland	0.64	0.22	0.22	0.12	0.94	0.21	0.50	0.12	0.81	0.21	0.27	0.13
Russia	0.13	0.21	0.03	0.12	0.40	0.20	0.09	0.12	-0.06	0.21	-0.21	0.12
Senegal	0.45	0.21	0.34	0.12	0.76	0.20	0.36	0.12	1.32	0.22	0.68	0.13
South-Africa	1.06	0.22	0.41	0.12	0.39	0.20	0.33	0.12	0.48	0.21	0.44	0.13
South-Korea	-1.97	0.21	-0.30	0.12	-0.92	0.20	0.18	0.12	-1.12	0.20	0.11	0.13
Sweden	0.26	0.21	0.16	0.12	0.45	0.20	0.34	0.12	0.17	0.20	0.16	0.12
Switzerland	-0.14	0.20	-0.10	0.12	-0.05	0.19	-0.01	0.11	0.48	0.20	0.13	0.12
USA	1.02	0.18	0.40	0.10	0.47	0.16	0.33	0.10	0.29	0.17	0.11	0.10
Being female	0.00	0.08	0.08	0.04	-0.30	0.07	-0.09	0.04	-0.32	0.08	-0.12	0.05
Age	0.02	0.01	0.01	0.00	0.03	0.01	0.01	0.00	0.03	0.01	0.02	0.00

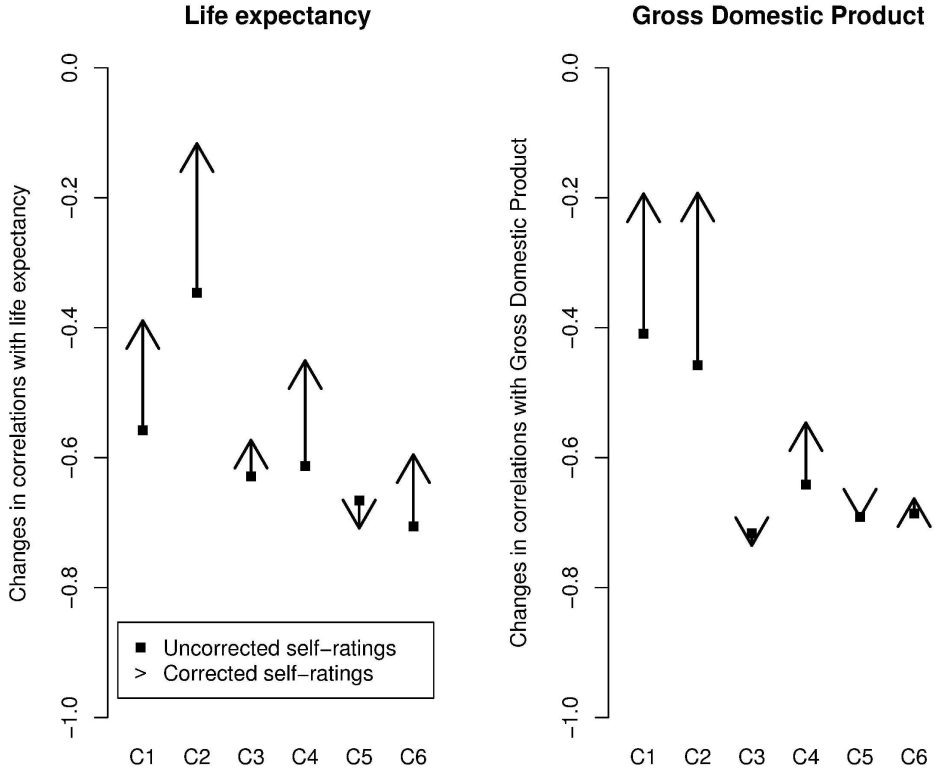
NOTE: Estim = Unstandardized regression coefficient; St Err = Standard error of regression coefficient; SOP = standard ordered probit model; COP = censored ordered probit model. Australia is the reference sample.

Table 4 (Continued). SOP and COP Regression Parameter Estimates and Standard Errors of the Estimates for the Six Facets of Conscientiousness.

	Achievement Striving				Self-Discipline				Deliberation			
	SOP		COP		SOP		COP		SOP		COP	
	Estim	St Err	Estim	St Err	Estim	St Err	Estim	St Err	Estim	St Err	Estim	St Err
Benin	2.35	0.22	0.77	0.12	1.88	0.21	1.32	0.13	1.56	0.22	0.68	0.59
Burkina Faso	2.15	0.23	0.76	0.13	1.64	0.22	1.13	0.12	1.41	0.22	0.76	0.03
China (Changchun)	0.82	0.20	0.38	0.12	1.02	0.20	0.61	0.12	0.53	0.20	0.47	0.38
China (Beijing)	0.35	0.18	0.22	0.10	1.50	0.18	0.16	0.10	0.45	0.17	0.42	-0.36
Estonia	0.02	0.19	0.09	0.11	0.13	0.19	0.14	0.11	0.58	0.19	0.29	0.12
Germany	0.42	0.23	0.19	0.14	0.70	0.23	0.10	0.13	-0.28	0.22	-0.01	0.10
Hong Kong	1.04	0.18	0.82	0.10	0.04	0.16	0.55	0.10	0.83	0.17	0.70	-0.55
Japan	-0.66	0.21	-0.25	0.12	-0.46	0.19	-0.41	0.11	-0.57	0.20	-0.02	-1.18
Lithuania	-0.53	0.18	-0.18	0.11	0.13	0.18	-0.32	0.11	-0.27	0.18	-0.13	-0.27
Malaysia	1.03	0.16	0.50	0.09	0.32	0.15	0.61	0.09	0.30	0.15	0.37	-0.41
Mali	2.23	0.24	0.60	0.13	1.82	0.23	1.29	0.13	1.18	0.22	0.62	-0.13
Mauritius	0.45	0.20	0.30	0.12	0.85	0.20	0.41	0.12	0.21	0.20	0.31	0.06
Philippines	1.37	0.19	0.57	0.11	1.02	0.18	0.65	0.11	0.49	0.18	0.35	0.01
Poland	0.44	0.20	0.37	0.12	0.60	0.21	0.09	0.12	-0.01	0.20	0.17	0.22
Russia	0.39	0.21	-0.09	0.12	0.16	0.20	0.22	0.12	0.34	0.20	0.23	0.03
Senegal	2.10	0.21	0.72	0.12	1.80	0.21	1.40	0.12	0.97	0.21	0.55	0.34
South-Africa	1.09	0.20	0.52	0.12	1.15	0.21	0.76	0.12	0.83	0.21	0.59	0.41
South-Korea	-1.28	0.21	0.11	0.12	-1.44	0.20	-0.42	0.12	-0.89	0.20	0.19	-0.30
Sweden	0.46	0.20	0.36	0.12	1.16	0.20	0.19	0.12	-0.47	0.20	0.05	0.16
Switzerland	0.02	0.20	0.07	0.11	0.59	0.20	-0.10	0.11	-0.18	0.20	0.07	-0.10
USA	0.42	0.17	0.13	0.10	0.84	0.17	0.30	0.10	0.09	0.16	0.13	0.40
Being female	-0.32	0.08	-0.14	0.04	-0.14	0.08	-0.12	0.04	0.06	0.08	0.08	0.08
Age	0.01	0.01	0.01	0.00	0.02	0.01	0.00	0.00	0.02	0.01	0.01	0.01

Figure Captions

Figure 1. Rank-order correlations of the uncorrected and corrected rankings of samples on the facets of Conscientiousness with culture-level objective criteria. C1 = Competence, C2 = Order, C3 = Dutifulness, C4 = Achievement Striving, C5 = Self-Discipline, C6 = Deliberation.



Rank-order correlations of the uncorrected and corrected rankings of samples on the facets of Conscientiousness with culture-level objective criteria. C1 = Competence, C2 = Order, C3 = Dutifulness, C4 = Achievement Striving, C5 = Self-Discipline, C6 = Deliberation.
197x174mm (600 x 600 DPI)