Are Cross-Cultural Comparisons of Personality Profiles Meaningful? Differential Item and Facet Functioning in the Revised NEO Personality Inventory

A. Timothy Church, Juan M. Alvarez, Nhu T. Q. Mai, Brian F. French, and Marcia S. Katigbak
Washington State University

Fernando A. Ortiz
Gonzaga University

Measurement invariance is a prerequisite for confident cross-cultural comparisons of personality profiles. Multigroup confirmatory factor analysis was used to detect differential item functioning (DIF) in factor loadings and intercepts for the Revised NEO Personality Inventory (P. T. Costa, Jr., & R. R. McCrae, 1992) in comparisons of college students in the United States ($N = 261$), Philippines ($N = 268$), and Mexico ($N = 775$). About 40%–50% of the items exhibited some form of DIF and item-level noninvariance often carried forward to the facet level at which scores are compared. After excluding DIF items, some facet scales were too short or unreliable for cross-cultural comparisons, and for some other facets, cultural mean differences were reduced or eliminated. The results indicate that considerable caution is warranted in cross-cultural comparisons of personality profiles.

Keywords: measurement invariance, differential item functioning, cross-cultural comparisons, Revised NEO Personality Inventory

The cross-cultural generalizability of the five-factor model (FFM) of personality has been demonstrated in many cultures (McCrae & Allik, 2002), particularly when measured by imported inventories such as the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992). Researchers have begun to draw conclusions about cultural differences in average trait levels by comparing mean profiles on the NEO-PI-R and other measures of the Big Five dimensions of Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness (McCrae, 2001, 2002; McCrae et al., 2010; Schmitt, Allik, McCrae, & Benet-Martínez, 2007; Terracciano et al., 2005). In the present study, we investigated the validity or meaningfulness of such comparisons by testing an important prerequisite, the cross-cultural measurement invariance of the NEO-PI-R at the item and facet levels.

We also address an important methodological issue—the extent to which lack of measurement invariance at the item level can carry forward to the scale level of personality inventories (French, Maller, & Zumbo, 2007; Li & Zumbo, 2009; Waller, Thompson, & Wenk, 2000; Zumbo, 2003; Zumbo & Koh, 2005). Although this issue has been addressed primarily by quantitative psychologists, it is also of central importance to personality psychologists, because it is at the scale level that scores on personality inventories are compared across cultures. In the case of the NEO-PI-R, lack of measurement invariance at the item level would call into question cross-cultural comparisons of facet scores. Lack of measurement invariance at the facet level would suggest that the facet scores are not equivalent indicators of the respective Big Five traits, and would call into question comparisons of either the facet or domain scores.

There are cogent theoretical reasons for comparing mean profiles across cultures. Such comparisons could increase our understanding of ecological, cultural, and biological influences on personality. For example, Hofstede and McCrae (2004) found relationships between dimensions of culture (i.e., Individualism, Uncertainty Avoidance, Power Distance, and Masculinity) and country-level means for the NEO-PI-R Big Five domain scores that suggested cultural influences on personality. Other researchers have observed higher average Extraversion and Openness to Experience scores of immigrant groups—or Europeans and Americans, as compared with Africans and Asians—and attributed these differences to selective emigration and resulting gene flow (Camperio Ciani, Capiluppi, Veronese, & Sartori, 2007; Olson, 2007). From an applied perspective, knowledge of cultural mean differences in trait levels, if large and reliable, might facilitate intercultural communications and adjustment.

This article was published Online First September 12, 2011.

A. Timothy Church, Juan M. Alvarez, Nhu T. Q. Mai, Brian F. French, and Marcia S. Katigbak, Department of Educational Leadership and Counseling Psychology, Washington State University; Fernando A. Ortiz, Counseling Center, Gonzaga University.

The research was supported by National Institute of Mental Health Grant MH59941 and National Science Foundation Grant 0953940. The data sets that were combined for this study were from previous studies that addressed different research questions and did not investigate differential item functioning (Church et al., 2007, 2008; Ortiz et al., 2007).

Correspondence concerning this article should be addressed to A. Timothy Church, Department of Educational Leadership and Counseling Psychology, Cleveland Hall, Washington State University, Pullman, WA 99164-2136. E-mail: church@mail.wsu.edu

Journal of Personality and Social Psychology, 2011, Vol. 101, No. 5, 1068–1089
 0022-3514/11/$12.00 DOI: 10.1037/a0025290

Despite their potential value, comparisons of mean profiles across cultures continue to generate controversy (e.g., Church, 2008; Poortinga, Van de Vijver, & Van Hemert, 2002). For example, Bock (2000) pointed out that one of the primary reasons for the downfall of the classic culture-and-personality school in anthropology was its tendency to characterize the personality or "national character" of whole populations, while deemphasizing individual variability. Bock has warned against fostering the same "uniformity assumption" in present cross-cultural studies of personality profiles. With this in mind, McCrae (2004) reminded readers that ascribing the mean trait level to all or specific individuals represents unwarranted stereotyping, because there is substantial within-culture variation in all traits. In addition, observed cultural differences are subject to a variety of substantive and artifactual interpretations that are difficult to disentangle, contributing further to questions about the validity of cross-cultural profile comparisons.

Evidence For and Against Mean Profile Comparisons

Several sources of evidence suggest that aggregate personality profiles may be meaningful. Perhaps the most persuasive is the geographical patterning of such profiles (Allik & McCrae, 2004; McCrae et al., 2010; Schmitt et al., 2007). For example, Allik and McCrae (2004) found that neighboring countries were generally grouped together in cluster and multidimensional scaling analyses of NEO-PI-R mean profiles. For example, European and American cultures were generally contrasted with Asian and African cultures, with the former higher in Extraversion and Openness to Experience and lower in Agreeableness. There were also a number of anomalies in the grouping of cultures, however. Schmitt et al. (2007) also reported sensible geographical patterning (with some anomalies) of Big Five Inventory profiles (Benet-Martínez & John, 1998).

Meaningful external correlates of country-level trait means have also been reported, including, for example, cultural dimensions, values, subjective well-being, organizational commitment, sociosexuality, and various health-related behaviors and outcomes (e.g., Allik & McCrae, 2004; Gelade, Dobson, & Gilbert, 2006; Hofstede & McCrae, 2004; McCrae & Terracciano, 2008; Schmitt et al., 2007). However, some correlations have been more difficult to interpret or have not replicated well across studies or methods. For example, McCrae and Terracciano (2008) identified 530 statistically significant ($p < .05$) culture-level correlates of NEO-PI-R scores in observer data, but only 272 of the relationships replicated in self-report data (McCrae, 2002).

Finally, the generalizability of country-level means across gender and age groups in both self-report and observer data suggests that these means may be reasonably valid (Costa, Terracciano, & McCrae, 2001; McCrae, 2001, 2002; McCrae, Terracciano, & 78 Members of the Personality Profiles of Cultures Project, 2005a; McCrae et al., 2010; Schmitt et al., 2007). The average cultural differences that are observed with personality inventories—whatever their meaning—are reliably found across gender and age.

However, the validity of cultural mean profiles may be called into question by the limited convergence of country-level means obtained with different Big Five measures (McCrae et al., 2010; Schmitt et al., 2007) and the failure of cultural mean profiles to converge with informants' ratings of typical personality in a culture (Church & Katigbak, 2002; McCrae et al., 2010; Terracciano et al., 2005)—what Terracciano et al. have referred to as "national character stereotypes." In addition, cross-cultural psychologists have noted various method and item biases that can reduce the measurement equivalence of personality inventories across cultures (Church, 2010; Vanderberg & Lance, 2000; van de Vijver & Leung, 1997). These include sampling biases, translation inequivalencies, cultural differences in response styles or negative item bias, differential familiarity with test materials, and items that are not equally relevant indicators of the personality constructs across cultures. Heine, Lehman, Peng, and Greenholtz (2002) have also argued that reference group effects—the tendency for respondents in different countries to rate their traits in comparison to different cultural norms or reference groups—will also confound cross-cultural mean comparisons.

Measurement Invariance

In the present study, we focused on one of the most fundamental factors that could invalidate—or at least raise significant concerns about—mean profile comparisons across cultures, a lack of measurement invariance. Researchers have delineated several levels of measurement invariance (Church, 2010; Vandenberg & Lance, 2000; van de Vijver & Leung, 1997). *Configural invariance* is exhibited when the same number of constructs or factors, and the same pattern of salient loadings, defines the structure of the instrument across cultures. *Metric invariance* is present when the factor loadings (i.e., regression slopes) for items that define the construct can be considered equal across cultures. Metric equivalence implies equivalent scale intervals, which facilitates comparisons of the nomological networks of the constructs across cultures (Steenkamp & Baumgartner, 1998). Meaningful comparisons of mean scores across cultures also require *scalar invariance*, a more stringent level in which the item intercepts can also be constrained to be equal across cultures. The item intercepts, which are the values of each item corresponding to the zero value of the latent construct or trait, indicate whether the measurement scales have the same origin or zero point across cultures (G. W. Cheung & Rensvold, 2000). Lack of measurement invariance at the item level is often referred to as *differential item functioning* (DIF).

Configural and metric invariance have already been demonstrated at the facet level for the NEO-PI-R using principal components analysis (e.g., Katigbak, Church, Guanzon-Lapeña, Carlota, & del Pilar, 2002; McCrae & Allik, 2002; Ortiz et al., 2007). However, *scalar invariance* is rarely tested in personality inventories. In particular, we could identify only two previous studies in which DIF was examined across cultures in measures of the Big Five dimensions. Nye, Roberts, Saucier, and Zhou (2008) found that over 60% of the items in the Big Five Mini-Markers (Saucier, 1994) exhibited DIF across American, Greek, and Chinese samples. Elimination of DIF items had dramatic effects on the comparisons of cultural means. Huang, Church, and Katigbak (1997) found that about 40% of the items in the original NEO-PI exhibited DIF and that many apparent cultural differences between Americans and Filipinos were eliminated after DIF items were removed. In a comparison of Americans and Germans, Johnson, Spinath, Krueger, Angleitner, and Riemann (2008) found that over one third of the items in the Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008) exhibited DIF and that

removal of DIF items eliminated cultural mean differences for all but one of the 11 scales. Applications of DIF methods to other personality measures have yielded variable results, with some researchers reporting small proportions of DIF items (i.e., 15% or less; Ellis, Becker, & Kimmel, 1993) and others reporting larger percentages (e.g., 20%–60%; Butcher, 1996; Taylor & Boeyens, 1991; van Leest, 1997; Waller et al., 2000).

On the basis of their MPQ results, Johnson et al. (2008) concluded that "it is not unreasonable to expect that many, if not most, of the translated questionnaires used in [cross-cultural] NEO-PI-R comparisons contain items that would show DIF if the appropriate analyses were carried out" (p. 693). If so, then some apparent cultural differences might be artifactual, whereas some valid cultural differences might be masked by DIF. In addition, when many DIF items are removed, it raises questions about construct underrepresentation and content validity in what remains of the instrument, making it unclear what aspects of the construct are being compared. Researchers differ in their views regarding the proportion of items that need to demonstrate invariance to enable cross-cultural comparisons (e.g., Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Ideally, most of the items will exhibit invariant factor loadings and intercepts, so that estimates of cultural means will be based largely on equivalent items.

Item Versus Scale-Level Invariance

In considering the measurement invariance issue, McCrae, Terracciano, & 79 Members of the Personality Profiles of Cultures Project (2005b) and McCrae and Terracciano (2008) suggested that DIF may cancel out at the scale level, allowing for valid cross-cultural comparisons of NEO-PI-R facet and domain scores. More generally, Labouvie and Ruetsch (1995) have argued that it is unrealistic to require invariance at the item level—for example, because most items are factorially complex—and that group comparisons require only invariance at the scale level. A few researchers have investigated the possibility that DIF might cancel out at the scale level in personality inventories.

Consistent with the suggestion of McCrae, Terracciano, & 79 Members (2005b), Waller et al. (2000) found substantial DIF in an investigation of Minnesota Multiphasic Personality Inventory factor scales, but little differential test functioning (DTF). Similarly, Reise, Smith, and Furr (2001) found that one third of the items in the NEO-PI-R Neuroticism scale exhibited gender DIF in a comparison of American men and women, but very little differential scale functioning was found. In contrast, Ellis and Mead (2000) found that about three fourths of the scales in a Spanish translation of the Sixteen Personality Factor Questionnaire (Cattell, Cattell, & Cattell, 1993) showed significant DTF, suggesting that DIF did translate to the scale level. Recent simulation studies suggest that DIF can influence the psychometric properties of test scores (e.g., coefficient alphas, score variances) depending on DIF type and severity (French & Finch, 2008; French & Maller, 2006) and, ultimately, statistical tests regarding group mean differences (Li & Zumbo, 2009). In short, more attention needs to be placed on examining the influence of DIF at the score level, the critical level at which decisions and interpretations about groups are typically made. Overall, previous findings with both real and simulated data, although not extensive, suggest that DIF may not cancel out at the scale or score level. However, this question has not yet been examined with the inventory that is most widely used in cross-national comparisons, the NEO-PI-R.

DIF Paradox

There is a paradox in the investigation of DIF. On the one hand, researchers interested in cross-cultural comparisons will prefer to find little or no DIF. On the other hand, items that exhibit DIF might reveal interesting cultural differences in the relevance or prevalence of the behavioral indicators of traits (Church, Katigbak, Miramontes, del Prado, & Cabrera, 2007; Huang et al., 1997; Johnson et al., 2008). Although a number of researchers have commented on the difficulty of explaining DIF, they nonetheless note the potential value of looking for patterns in the content of DIF items (Ellis, 1990; Ellis et al., 1993; Huang et al., 1997; Johnson et al., 2008; Reise et al., 2001). Some DIF could be the result of translation inequivalencies, sampling differences, motivational differences, or how cultural groups respond to particular item formats (e.g., reverse-keyed items). Of greater interest for personality psychologists would be cultural differences in the relevance of the behaviors (items) as indicators of the associated traits, or the prevalence of the behaviors in the cultural groups. Researchers have tried to explain such differences in terms of cultural values, norms, or practices. This DIF paradox reminds us that some DIF may represent valid cultural differences in the behavioral indicators of traits, rather than measurement artifacts. Even so, the presence of DIF detracts from measurement invariance and researchers' ability to make confident mean comparisons across cultures.

Overview of the Present Study

We had several goals. The primary goal was to determine the meaningfulness of cross-cultural comparisons of mean profiles by examining the item- and facet-level invariance of the NEO-PI-R across three cultural samples, the United States, Mexico, and the Philippines. Second, we sought to address the methodological question of whether lack of invariance at the item level carries forward to the facet level at which personality scores are actually compared across cultures. In so doing, we also tested McCrae, Terracciano, & 79 Members's (2005b) proposal that DIF may cancel out at the facet level. Third, we wished to determine whether elimination of DIF items impacts apparent cultural mean differences in trait levels. Fourth, we sought to explore whether DIF items can reveal meaningful patterns of cultural differences in the behavioral indicators of personality traits.

We used multigroup confirmatory factor analysis (CFA) to achieve these goals. CFA was favored over item response theory (IRT) methods (Camilli & Shepard, 1994; Embretson & Reise, 2000) for several reasons. First, use of CFA enabled us to test for item-level and facet-level invariance using the same analytic strategy. Second, whereas our sample sizes were sufficient for CFA (Kahn, 2006), only one of our three cultural samples met the larger sample size requirement (e.g., 500 or more) typically recommended for IRT methods (Stark, Chernyshenko, & Drasgow, 2006). Third, IRT methods are more complex than CFA and thus less accessible to most readers, as well as researchers who may want to conduct their own measurement invariance analyses. For

detailed comparisons of CFA and IRT-based DIF analyses, see Stark et al. (2006); Raju, Laffitte, and Byrne (2002); Reise, Widaman, and Pugh (1993); Meade and Lautenschlager (2004); and Finch and French (2007).

We investigated DIF for each pairwise comparison among the three cultures, that is, the United States versus the Philippines, the United States versus Mexico, and the Philippines versus Mexico. This provided an opportunity to replicate DIF across comparisons to determine whether some items are more prone to DIF than others, or whether DIF tends to be specific to particular cultural comparisons. In each pairwise cultural comparison, we expected to find a moderate proportion of DIF items (e.g., 30%–40%) in the NEO-PI-R and that at least some DIF would carry forward to the facet or score level. We also expected some cultural mean differences in NEO-PI-R facet scores to be impacted by elimination of DIF items. If so, it would suggest that some cultural mean comparisons can be distorted by the presence of DIF.

Method

Sample.

The United States. The United States sample included 261 students (69 men, 180 women, 12 not reporting gender) at Washington State University, a midsize public university. Mean age was 21.48 (SD = 3.60). Self-reported ethnic backgrounds were as follows: White/Caucasian (81.6%), Chicano/Latino/Hispanic (2.7%), Asian/Pacific Islander (2.3%), African American (1.5%), Native American (.4%), bi- or multiracial (5.0%), and "other" or not reporting (6.5%). Students received extra credit in their courses for participation. This sample was obtained by combining the participants from two previous studies that related NEO-PI-R domain scores to daily behaviors but did not examine DIF (Church et al., 2007, 2008).

The Philippines. The Philippine sample included 268 students (103 men, 149 women, 16 not reporting gender) at two private universities in Manila (University of Santo Tomas, n = 168; De La Salle University, n = 80), and a smaller college in a medium-size city 60 km south of Manila (De La Salle Lipa, n = 20). Mean age was 18.40 (SD = 1.45). About 94% self-reported their ethnicity as Filipino, whereas the remaining participants reported Chinese (4%) or biracial (e.g., Filipino-Chinese) ethnicity (2%). The sample was obtained by combining the participants from two previous studies that related NEO-PI-R domain scores to daily behaviors but did not examine DIF (Church et al., 2007, 2008).

Mexico. The Mexican sample included 775 students (302 men, 473 women) from the National Autonomous University of Mexico at Iztacala (n = 202), the Hidalgan Institute of Higher Learning Studies (n = 189), and the Autonomous University of Yucatan (n = 384). Mean age was 19.79 (SD = 2.41). All participants reported their ethnic background to be Mestizo, the predominant ethnicity in Mexico and a mixture of European (usually Spanish) and American Indian ancestry. The sample was previously used in a study in which the relationship between the FFM and indigenous Mexican personality inventories were investigated, but it did not examine DIF (Ortiz et al., 2007).

The NEO-PI-R. The 240-item NEO-PI-R (Costa & McCrae, 1992) measures the Big Five traits of Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. Each domain encompasses six eight-item facet scales that measure specific traits in each Big Five domain (see Table 1). Items were rated on a 5-point scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, 5 = *strongly agree*). The NEO-PI-R has been translated in over 70 cultures (McCrae & Allik, 2002). In the Philippine sample, the Filipino (Tagalog) version was used, which was previously derived using back-translation methods (Del Pilar, 1998; McCrae, Costa, del Pilar, Rolland, & Parker, 1998). Previous studies have demonstrated the generalizability of the facet-level structure of the NEO-PI-R to Philippine samples and the convergent validity of the Filipino NEO-PI-R with indigenous trait and behavior measures (Church et al., 2007, 2008; Katigbak et al., 2002; McCrae et al., 1998). In discussing the Filipino translation, McCrae et al. (1998) noted that four items were replaced entirely with content deemed to be more culturally relevant. In addition, our inspection of the Filipino translation revealed six more items that were viewed as significant modifications or substitutions, plus 12 items (including one of the modified items) that reversed the keying of the original English language item (the reverse-keyed items were, or course, recoded appropriately before conducting the DIF analyses). Rather than discard these items, they were retained

Table 1
Internal Consistency Reliabilities (αs) for NEO-PI-R Facets in Three Cultures

NEO-PI-R facet	United States	Philippines	Mexico
Neuroticism			
N1: Anxiety	.72	.64	.57
N2: Hostility	.69	.76	.70
N3: Depression	.80	.77	.69
N4: Self-consciousness	.66	.58	.62
N5: Impulsivity	.63	.44	.55
N6: Vulnerability	.70	.67	.68
Extraversion			
E1: Warmth	.73	.68	.66
E2: Gregariousness	.77	.73	.67
E3: Assertiveness	.76	.72	.68
E4: Activity	.61	.57	.55
E5: Excitement-Seeking	.63	.49	.50
E6: Positive emotions	.77	.64	.70
Openness to Experience			
O1: Fantasy	.76	.62	.62
O2: Aesthetics	.79	.58	.69
O3: Feelings	.71	.64	.52
O4: Actions	.55	.51	.24
O5: Ideas	.79	.66	.75
O6: Values	.69	.46	.31
Agreeableness			
A1: Trust	.82	.68	.70
A2: Straightforwardness	.67	.61	.61
A3: Altruism	.74	.57	.63
A4: Compliance	.68	.63	.55
A5: Modesty	.70	.66	.65
A6: Tender-Mindedness	.58	.57	.32
Conscientiousness			
C1: Competence	.65	.66	.62
C2: Order	.68	.71	.46
C3: Dutifulness	.64	.52	.61
C4: Achievement-Striving	.73	.71	.62
C5: Self-Discipline	.81	.75	.74
C6: Deliberation	.70	.70	.69

Note. NEO-PI-R = Revised NEO Personality Inventory.

because the present DIF analyses would enable us to test whether these substituted or reversed items provided equivalent measurement, as compared with the original items.

In Mexico, a Spanish version of the NEO-PI-R (Gellman, 1994) was used. Gellman (1994) reported acceptable equivalence of the English and Spanish versions in a bilingual test–retest study with college students in the United States, and the Spanish version has also been used previously in Mexico (McCrae, Terracciano, & 78 Members, 2005a). Ortiz et al. (2007) had it further reviewed by a clinical psychology professor in Mexico and a Spanish language professor at Washington State University, both of whom were born in Mexico and fluent in Mexican Spanish. On the basis of their recommendations, some minor corrections in grammar and syntax were made. Our inspection of the Spanish translation revealed only one item that qualifies as a culture-specific substitution. The E5: Excitement-seeking item about not wishing to vacation in Las Vegas refers instead to Cancun. No Spanish translations resulted in a reversal of the keying of the original items. Ortiz et al. (2007) demonstrated the generalizability of the facet-level structure of the NEO-PI-R and the convergent validity of the NEO-PI-R domain scores with indigenous Mexican inventories.

Table 1 shows the internal consistency (α) reliabilities of the 30 facet scales in each sample. The α values provide one indication of measurement invariance at the facet level. The mean αs across all 30 facets were .71 in the United States (*Mdn* = .70), .61 in the Philippines (*Mdn* = .64), and .60 in Mexico (*Mdn* = .62). A few facets had α values lower than .50 in the Filipino and Mexican samples, which are clearly marginal (Nunnally & Bernstein, 1994).

Procedure. In the United States, some students were paid for research participation and completed the NEO-PI-R in monitored groups, whereas others received extra credit in class and completed the instrument on their own time. In the Philippines, some students were paid and completed the NEO-PI-R in monitored groups, whereas others volunteered to complete the instrument in regular classes. In Mexico, volunteer students completed the NEO-PI-R during regular classes.

Data analysis. The samples described above excluded a small number of respondents in each culture who had left 12 or more NEO-PI-R items blank. Following the procedure described in the NEO-PI-R manual (Costa & McCrae, 1992), any remaining blank responses were replaced with a neutral response.

Multigroup CFA was used, as implemented by the AMOS 16.0 program, to test the measurement invariance of the NEO-PI-R items, then the facets, in pairwise comparisons of the three cultural samples. The extent of DIF was examined in each of the 30 NEO-PI-R facet scales, as well as the extent to which DIF canceled out or carried forward to the facet level. Maximum likelihood estimation was used to derive all model parameters.

In a CFA framework, the raw score for an item can be expressed as shown in Equation 1, where the latent variable is the facet-level trait (e.g., N1: Anxiety). Scalar invariance holds when the regression parameters (i.e., intercept and loading) can be considered equal across groups, so that people in two different cultures who have the same position on the latent variable have the same raw score.

$$\text{Equation 1: Raw item score} = \text{Intercept} + \text{Loading (Latent Variable)}.$$

To conduct DIF analyses, it is first necessary to select one of the eight items in each facet to serve as a referent or anchor item to identify the CFA model and define the metric of the latent variable or facet. To do so, a principal components analysis was conducted on the eight items in each facet in each culture. For each cultural comparison (e.g., United States vs. Philippines), what was selected as the anchor was the item within each facet that had the most similar factor loadings in the two cultural samples (Finch & French, 2008). Across the three cultural comparisons, the mean cultural difference in the factor loadings for the anchor items was .03, and the factor loadings for the anchor items ranged from .30 to .83. These principal components analyses also served to demonstrate the essential unidimensionality of the items in each facet scale. Thus, although responses to personality items may be more complex or multifaceted than mental ability items, it is appropriate to conduct DIF analyses on them (Reise et al., 2001).

CFA can identify DIF in the item factor loadings or intercepts. Assuming translation equivalence, loading DIF can indicate that the thought, feeling, or behavior referred to in the item is less *relevant* as an indicator of the trait in one of the two cultures being compared (and conversely, more relevant in the other culture). If an item exhibits intercept DIF, it can indicate that the thought, feeling, or behavior is less *prevalent* or endorsed less in one of the two cultures by individuals with comparable levels of the trait. It is only meaningful to test for intercept DIF if the factor loading (i.e., regression slope) for a given item is invariant across cultures.[1] Accordingly, the following steps were used to examine the measurement invariance of the items in each facet for each cultural comparison (Steenkamp & Baumgarten, 1998). The latent variable in each DIF analysis was one of the 30 facets in the NEO-PI-R (e.g., N1: Anxiety), and the eight items in the facet scale were the observed indicators.

Testing for factor loading DIF:

Step 1. Test a CFA model in which the factor loadings for all items in the facet are freely estimated in both cultures (except for the anchor item, which is fixed to 1.0 in each culture).

Step 2. Test a CFA model in which all of the item factor loadings are constrained to be equal in the two cultures.

[1] Recently, Stark et al. (2006) offered several reasons for testing DIF in factor loadings and intercepts simultaneously: to make the analysis less cumbersome, to reduce the total number of statistical tests and Type I error, and to avoid errors in DIF detection for the loadings propagating to subsequent tests of intercept DIF. Also, because item slopes (loadings) and intercepts can be correlated to some extent, metric (loading) and scalar (intercept) invariance are not necessarily strictly independent. Finally, in practice, researchers are likely to discard items that exhibit either kind of DIF, so it is arguably unimportant whether an item exhibits DIF in the loading or intercept. However, if one is interested, as we were, in differentiating the amount of loading and intercept DIF—because the two types of DIF have different conceptual meanings and implications for cross-cultural measurement—it is necessary to test for the two types of DIF separately. Indeed, this is the common practice in DIF analyses based on CFA (Steenkamp & Baumgarten, 1998; Vandenberg & Lance, 2000).

Step 3. Compare the overall fit of the constrained (Step 2) and freely estimated (Step 1) models using a chi-square difference test. If the chi-square difference test is not statistically significant ($p < .01$), then there is no DIF in the item factor loadings for that facet, and one should proceed to test for DIF in the item intercepts (i.e., Step 6). If the chi-square difference test is statistically significant, then proceed to Step 4 to test for factor loading DIF in the individual items.

Step 4. Successively test for factor loading DIF in each individual (focal) item by constraining the factor loadings for that item to be equal in the two cultures, while freely estimating the factor loadings of all the remaining items (except the anchor item) in both cultures.

Step 5. Compare the fit of each of the models in Step 4, in which the factor loading for a focal item is constrained to be equal in the two cultures, with the model in Step 1 in which all factor loadings were freely estimated in the two cultures. If the chi-square difference test is statistically significant, then the factor loading for the focal item exhibits DIF.

Testing for intercept DIF:

Step 6. Test a CFA model in which all item intercepts, including for the anchor item, are freely estimated in the two cultures. In this model, and in all subsequent models testing for intercept DIF, retain the factor loading equality constraints for those items that did not exhibit factor loading DIF in Step 5.

Step 7. Test a CFA model in which all item intercepts are constrained to be equal in the two cultures.

Step 8. Compare the fit of the constrained (Step 7) and freely estimated (Step 6) intercept models using a chi-square difference test. If the chi-square difference test is not statistically significant ($p < .01$), then there is no DIF in the item intercepts for that facet. If the chi-square difference test is statistically significant, then proceed to Step 9 to test for intercept DIF in the individual items.

Step 9. Successively test for intercept DIF in each (focal) item that did not exhibit factor loading DIF. Do so by constraining the intercept for that item to be equal in the two cultures, while freely estimating the intercept of all the remaining items in both cultures.

Step 10. Compare the fit of each of the models in Step 9, in which the intercept for a focal item is constrained to be equal in the two cultures, with the model in Step 7, in which all intercepts were freely estimated in the two cultures. If the chi-square difference test is statistically significant, then the intercept for the focal item exhibits intercept DIF.

After these DIF analyses were completed, analogous CFA analyses were conducted at the facet level to test for measurement invariance in the factor loadings and intercepts of each facet as an indicator of the relevant Big Five dimension. For example, for the Neuroticism domain, Neuroticism was the single latent variable in the CFA model, and scores for the six facet scales were the observed indicators. These "higher level" facet analyses enabled us to determine the extent to which DIF within each facet scale canceled out or carried forward to the facet level at which NEO-PI-R scores are typically computed and compared.

In the above steps, the chi-square difference test provides a direct statistical test of whether the factor loadings or intercepts can be considered equivalent in two cultures. However, when sample sizes are large, even modest differences in loadings or intercepts can result in significant changes in chi-square values between constrained and freely estimated models. In addressing this issue of statistical versus practical significance, G. W. Cheung and Rensvold (2002) recommended that researchers consult Bentler's (1990) comparative fit index (CFI) to identify items lacking measurement invariance. The CFI is a relative fit index that quantifies the fit of each model in terms of its proportional improvement over a null model of no covariation between the observed variables (i.e., no common factors) (Bentler, 1990).[2] Specifically, G. W. Cheung and Rensvold recommended that changes in the CFI (ΔCFI) between constrained and freely estimated models that exceed .01 be used to designate items that lack measurement invariance. In a simulation study, these authors found that the ΔCFI value has a number of desirable features. It is independent of both model complexity and sample size and uncorrelated with overall model fit indices.

Accordingly, two criteria were used to designate items or facets as noninvariant. The chi-square difference test criterion ($\Delta\chi^2$)—which was applied in the sequence of steps outlined above—provides a direct test of statistical significance and is the most commonly applied criterion in the literature. Because of the number of statistical tests, a more conservative alpha level of .01 was used for the chi-square difference test criterion. The ΔCFI $> .01$ criterion has only recently been applied in DIF analyses (e.g., French & Finch, 2006). However, unlike the chi-square difference test, it is independent of sample size and thus will be more equivalent across our three pairwise cultural comparisons. We make use of both criteria in reporting the amount of DIF, but rely on the more conservative ΔCFI $> .01$ criterion for some analyses, because this criterion points to DIF of greater practical significance. In the present study, the mean loading differences for loading DIF items identified using the ΔCFI criteria were substantial ($M = 0.41$, $SD = 0.18$), and the mean intercept differences for intercept DIF items corresponded to about half a scale point ($M = 0.52$, $SD = 0.27$). It should be noted that any item that qualifies as a DIF item based on the ΔCFI $> .01$ criterion will also satisfy the chi-square difference test criterion.

[2] Bentler's (1990) CFI index is computed as $1 - (\max[\lambda_m,0]/\max[\lambda_n, \lambda_m, 0])$, where $\lambda_m = \chi^2_{model} - df_{model}$, $\lambda_n = \chi^2_{null} - df_{null}$. The index is normed to values between 0 and 1. Its rationale derives from the fact that the maximum likelihood fit function is distributed asymptotically as a noncentral chi-square distribution with noncentrality parameter λ when the hypothesized model is false.

Results

Overview. We address the following questions in the results: How much DIF is exhibited in the pairwise cultural comparisons? Are there consistencies across cultural comparisons in the items and facets that exhibit the most DIF? When DIF is detected, is there any pattern in which cultural samples exhibit the highest loadings or intercepts? Does DIF in the factor loadings or intercepts cancel out or carry forward to the facet level at which scores are compared across cultures? Are mean differences between cultures on the facet scales impacted by elimination of DIF items? Finally, can any patterns in the content of the DIF items be detected?

The DIF. Tables 2,3, and 4 summarize the results for comparisons of the United States and Philippines, United States and Mexico, and Philippines and Mexico, respectively. The first two columns in each table show the number of items in each facet scale exhibiting loading DIF based on the chi-square difference test ($\Delta\chi^2$) and ΔCFI $> .01$ criteria, respectively. In the third column of Tables 2–4, we show the mean ΔCFI values for those items in each facet for which loading DIF was detected on the basis of the chi-square difference tests. The ΔCFI values provide a measure of the size of the model misfit that results from constraining item loadings or intercepts to equality across cultures. In the fourth column of these tables, we show, for those items with loading DIF based on the chi-square difference tests, ratios showing the number of DIF items that had higher loadings in each of the two cultures when the loadings were freely estimated. These ratios indicate whether there were any consistent tendencies for one or the other culture to exhibit higher freely estimated loadings when the loadings could not be constrained to equality. The next several columns in each table contain analogous information related to intercept DIF.

Table 2
Summary of Noninvariant NEO-PI-R Items and Facets in U.S.-Philippines Comparison

	Factor loadings				Intercepts				Facet noninvariance			
	# of DIF items				# of DIF items				Loadings		Intercepts	
NEO-PI-R facet	$\Delta\chi^2$	ΔCFI	Mean ΔCFI	US:Phil. ratio	$\Delta\chi^2$	ΔCFI	Mean ΔCFI	US:Phil. ratio	$\Delta\chi^2$	ΔCFI	$\Delta\chi^2$	ΔCFI
Neuroticism												
N1: Anxiety	0	0	—	—	2	2	.05	1:1	7.13 U.S.	.00	—	—
N2: Angry Hostility	1	0	.00	0:1	4	4	.04	1:3	17.24 Phil.	.01	—	—
N3: Depression	0	0	—	—	2	2	.04	0:2	Anchor	—	7.72 Phil.	.01
N4: Self-Consciousness	0	0	—	—	5	4	.11	1:4	—	—	33.68 Phil.	.03
N5: Impulsivity	0	0	—	—	4	4	.14	3:1	—	—	23.70 U.S.	.02
N6: Vulnerability	0	0	—	—	2	2	.05	0:2	—	—	9.85 Phil.	.01
Extraversion												
E1: Warmth	0	0	—	—	6	5	.09	6:0	—	—	68.55 U.S.	.09
E2: Gregariousness	0	0	—	—	4	1	.02	2:2	Anchor	—	—	—
E3: Assertiveness	2	2	.02	2:0	4	2	.05	2:2	—	—	6.85 U.S.	.01
E4: Activity	0	0	—	—	3	3	.15	3:0	—	—	31.95 U.S.	.04
E5: Excitement-Seeking	0	0	—	—	4	4	.07	4:0	—	—	27.01 U.S.	.04
E6: Positive Emotions	2	0	.01	2:0	2	2	.06	2:0	—	—	42.11 U.S.	.05
Openness to Experience												
O1: Fantasy	1	0	.01	1:0	5	4	.04	4:1	—	—	13.70 U.S.	.03
O2: Aesthetics	0	0	—	—	6	6	.04	1:5	Anchor	—	11.79 Phil.	.02
O3: Feelings	0	0	—	—	4	4	.05	3:1	—	—	—	—
O4: Actions	0	0	—	—	4	4	.32	1:3	—	—	74.48 Phil.	.18
O5: Ideas	0	0	—	—	3	1	.02	0:3	—	—	9.38 Phil.	.02
O6: Values	0	0	—	—	7	7	.30	7:0	—	—	211.64 U.S.	.51
Agreeableness												
A1: Trust	2	1	.02	2:0	3	1	.02	2:1	—	—	34.48 U.S.	.05
A2: Straightforwardness	1	1	.05	0:1	5	5	.09	4:1	—	—	30.73 U.S.	.04
A3: Altruism	1	1	.02	0:1	3	3	.16	3:0	—	—	62.00 U.S.	.08
A4: Compliance	2	2	.03	1:1	5	5	.05	2:3	—	—	—	—
A5: Modesty	0	0	—	—	3	3	.04	3:0	Anchor	—	9.64 U.S.	.01
A6: Tender-Mindedness	1	1	.03	1:0	6	5	.21	2:4	—	—	17.78 Phil.	.02
Conscientiousness												
C1: Competence	0	0	—	—	5	5	.06	4:1	—	—	10.94 U.S.	.01
C2: Order	3	2	.02	1:2	5	1	.01	3:2	—	—	—	—
C3: Dutifulness	0	0	—	—	2	2	.05	1:1	Anchor	—	—	—
C4: Achievement-Striving	1	0	.01	1:0	2	1	.06	0:2	—	—	15.07 Phil.	.02
C5: Self-Discipline	0	0	—	—	4	2	.02	1:3	—	—	—	—
C6: Deliberation	0	0	—	—	4	3	.04	0:4	—	—	23.49 Phil.	.02

Note. In the Facet noninvariance columns, facets with significant loading or intercept noninvariance ($\Delta\chi^2$) are annotated to indicate the culture with the higher loading or intercept. NEO-PI-R = Revised NEO Personality Inventory; DIF = differential item functioning; CFI= comparative fit index; U.S. = United States; Phil. = Philippines. Dashes indicate the absence of DIF or facet noninvariance involving the loadings or intercepts for a given facet.

Table 3
Summary of Noninvariant NEO-PI-R Items and Facets in U.S.-Mexico Comparison

	Factor loadings				Intercepts				Facet noninvariance			
	# of DIF items				# of DIF items				Loadings		Intercepts	
NEO-PI-R facet	$\Delta\chi^2$	ΔCFI	Mean ΔCFI	US:Mex. ratio	$\Delta\chi^2$	ΔCFI	Mean ΔCFI	US:Mex. ratio	$\Delta\chi^2$	ΔCFI	$\Delta\chi^2$	ΔCFI
Neuroticism												
N1: Anxiety	4	3	.02	4:0	4	3	.03	3:1	—	—	26.13 U.S.	.01
N2: Angry Hostility	3	1	.01	1:2	2	2	.02	0:2	13.06 Mex.	.01	—	—
N3: Depression	4	1	.01	4:0	3	3	.04	3:0	—	—	16.28 U.S.	.01
N4: Self-Consciousness	0	0	—	—	5	5	.04	3:2	Anchor	—	23.23 U.S.	.01
N5: Impulsivity	1	1	.03	1:0	6	3	.03	5:1	—	—	36.00 U.S.	.02
N6: Vulnerability	0	0	—	—	5	3	.02	2:3	—	—	—	—
Extraversion												
E1: Warmth	0	0	—	—	6	5	.06	6:0	—	—	100.26 U.S.	.06
E2: Gregariousness	0	0	—	—	5	4	.02	2:3	—	—	—	—
E3: Assertiveness	1	0	.01	1:0	4	4	.05	2:2	—	—	—	—
E4: Activity	0	0	—	—	6	5	.04	5:1	—	—	22.01 U.S.	.01
E5: Excitement-Seeking	1	0	.01	1:0	6	3	.06	5:1	Anchor	—	—	—
E6: Positive Emotions	2	0	.00	1:1	2	0	.01	2:0	—	—	—	—
Openness to Experience												
O1: Fantasy	0	0	—	—	6	4	.02	4:2	—	—	—	—
O2: Aesthetics	0	0	—	—	7	4	.02	2:5	—	—	14.21 Mex.	.02
O3: Feelings	0	0	—	—	5	5	.08	5:0	Anchor	—	69.68 U.S.	.08
O4: Actions	5	5	.07	4:1	1	1	.12	0:1	—	—	300.53 Mex.	.33
O5: Ideas	1	0	.00	0:1	6	1	.01	1:5	—	—	16.33 Mex.	.02
O6: Values	3	3	.07	3:0	3	3	.09	1:2	—	—	—	—
Agreeableness												
A1: Trust	1	0	.01	1:0	7	4	.02	7:0	Anchor	—	141.63 U.S.	.03
A2: Straightforwardness	0	0	—	—	4	3	.05	4:0	—	—	67.64 U.S.	.06
A3: Altruism	0	0	—	—	7	4	.03	7:0	—	—	58.89 U.S.	.05
A4: Compliance	1	0	.01	1:0	3	2	.04	2:1	—	—	27.94 U.S.	.02
A5: Modesty	2	1	.02	1:1	6	6	.05	6:0	—	—	92.53 U.S.	.08
A6: Tender-Mindedness	1	1	.07	1:0	7	7	.11	2:5	—	—	16.78 U.S.	.01
Conscientiousness												
C1: Competence	0	0	—	—	3	0	.01	3:0	—	—	—	—
C2: Order	2	2	.02	2:0	2	1	.03	2:0	13.61 U.S.	.00	—	—
C3: Dutifulness	2	0	.01	2:0	1	1	.05	0:1	—	—	17.29 Mex.	.00
C4: Achievement-Striving	1	0	.01	1:0	6	3	.02	2:4	—	—	8.03 Mex.	.00
C5: Self-Discipline	1	0	.00	1:0	4	2	.02	2:2	Anchor	—	—	—
C6: Deliberation	0	0	—	—	4	3	.02	1:3	—	—	9.78 Mex.	.00

Note. In the Facet noninvariance columns, facets with significant loading or intercept noninvariance ($\Delta\chi^2$) are annotated to indicate the culture with the higher loading or intercept. NEO-PI-R = Revised NEO Personality Inventory; DIF = differential item functioning; CFI= comparative fit index; U.S. = United States; Mex. = Mexico. Dashes indicate the absence of DIF or facet noninvariance involving the loadings or intercepts for a given facet.

As an example in Table 2, only one item in the N2: Angry Hostility facet was identified as having loading DIF on the basis of the chi-square difference tests, and there were no loading DIF items based on the more conservative ΔCFI criterion. Although statistically significant in the chi-square difference test, the change in model fit resulting from constraining the factor loading for the DIF item was very small (mean ΔCFI = .00). The 0:1 ratio in the fourth column indicates that the freely estimated factor loading for the single loading DIF item was larger in the Philippine sample. Four of the items in the N2: Angry Hostility facet were identified as intercept DIF items on the basis of both the $\Delta\chi^2$ and ΔCFI criteria. Model misfit resulting from constraining the intercepts was moderately large (mean ΔCFI = .04). Finally, one intercept was larger in the United States sample, and three intercepts were larger in the Philippine sample.

Amount of loading and intercept DIF. One noteworthy finding in Tables 2, 3, and 4 was that loading DIF was relatively infrequent in all three pairwise cultural comparisons, whereas intercept DIF was quite frequent. This ratio of occurrence is consistent with what is seen in large standardized tests and has been used in DIF simulation work (e.g., French & Maller, 2007). Across the three cultural comparisons, the percentage of items that exhibited loading DIF ranged from 7.1% to 18.8% as indexed by the $\Delta\chi^2$ criterion and from 4.2% to 9.6% as indexed by the ΔCFI criterion. In contrast, the percentages of items that exhibited intercept DIF ranged from 47.5% to 56.7% as indexed by the $\Delta\chi^2$ criterion and from 39.1% to 40.4% as indexed by the ΔCFI criterion (and recall that intercept DIF is not tested for those items that exhibit loading DIF). In addition, as seen in the tables, the misfit in models (i.e., mean ΔCFI values) that resulted from constraining factor loadings was generally modest, whereas constraints on the item intercepts resulted in more substantial model misfit. In total, the percentages of items exhibiting some form of DIF in the three pairwise cultural comparisons ranged from 56.3%

Table 4
Summary of Noninvariant NEO-PI-R Items and Facets in Philippines-Mexico Comparison

	Factor loadings				Intercepts				Facet noninvariance			
	# of DIF items				# of DIF items				Loadings		Intercepts	
NEO-PI-R facet	$\Delta\chi^2$	ΔCFI	Mean ΔCFI	Phil.:Mex. ratio	$\Delta\chi^2$	ΔCFI	Mean ΔCFI	Phil.:Mex. ratio	$\Delta\chi^2$	ΔCFI	$\Delta\chi^2$	ΔCFI
Neuroticism												
N1: Anxiety	3	2	.02	2:1	4	3	.07	3:1	—	—	47.36 Phil.	.02
N2: Angry Hostility	3	1	.01	2:1	1	1	.02	0:1	Anchor	—	—	—
N3: Depression	3	1	.01	3:0	3	3	.07	3:0	—	—	53.45 Phil.	.02
N4: Self-Consciousness	0	0	—	—	7	7	.15	6.1	7.89 Mex.	.00	—	—
N5: Impulsivity	2	1	.02	2:0	5	5	.06	3:2	18.40 Mex.	.01	—	—
N6: Vulnerability	0	0	—	—	4	3	.03	3:1	—	—	10.42 Phil.	.00
Extraversion												
E1: Warmth	0	0	—	—	3	3	.06	2:1	Anchor	—	—	—
E2: Gregariousness	2	0	.00	1:1	2	1	.04	2:0	—	—	—	—
E3: Assertiveness	0	0	—	—	4	2	.03	0:4	—	—	—	—
E4: Activity	2	0	.01	2:0	6	4	.03	3:3	—	—	—	—
E5: Excitement-Seeking	3	3	.02	0:3	5	5	.09	2:3	—	—	18.51 Mex.	.01
E6: Positive Emotions	1	0	.01	0:1	2	2	.03	1:1	7.46 Mex.	.01	—	—
Openness to Experience												
O1: Fantasy	1	1	.02	1:1	4	1	.01	1:3	—	—	—	—
O2: Aesthetics	0	0	—	—	5	2	.03	2:3	—	—	—	—
O3: Feelings	0	0	—	—	4	4	.11	4:0	—	—	40.98 Phil.	.04
O4: Actions	3	3	.07	3:0	4	4	.12	1:3	—	—	100.32 Mex.	.11
O5: Ideas	0	0	—	—	1	0	.01	1:0	Anchor	—	—	—
O6: Values	3	3	.07	3:0	5	5	.27	0:5	—	—	251.28 Mex.	.29
Agreeableness												
A1: Trust	4	0	.01	2:2	2	2	.03	1:1	—	—	—	—
A2: Straightforwardness	3	1	.01	1:2	3	2	.02	1:2	—	—	—	—
A3: Altruism	0	0	—	—	4	4	.06	2:2	Anchor	—	—	—
A4: Compliance	1	0	.01	0:1	4	4	.10	4:0	21.68 Phil.	.02	—	—
A5: Modesty	0	0	—	—	5	5	.06	4:1	16.66 Phil.	.02	—	—
A6: Tender-Mindedness	4	1	.02	3:1	4	4	.24	2:2	—	—	83.02 Phil.	.08
Conscientiousness												
C1: Competence	0	0	—	—	5	4	.03	1:4	—	—	—	—
C2: Order	5	5	.03	5:0	3	2	.03	3:0	23.49 Phil.	.01	—	—
C3: Dutifulness	0	0	—	—	4	4	.05	1:3	—	—	20.22 Mex.	.01
C4: Achievement-Striving	1	1	.02	1:0	4	4	.06	2:2	—	—	—	—
C5: Self-Discipline	1	0	.00	0:1	2	1	.04	1:1	Anchor	—	—	—
C6: Deliberation	0	0	—	—	5	3	.03	4:1	—	—	7.16 Phil.	.00

Note. In the Facet noninvariance columns, facets with significant loading or intercept noninvariance ($\Delta\chi^2$) are annotated to indicate the culture with the higher loading or intercept. NEO-PI-R = Revised NEO Personality Inventory; DIF = differential item functioning; CFI= comparative fit index; Phil. = Philippines; Mex. = Mexico. Dashes indicate the absence of DIF or facet noninvariance involving the loadings or intercepts for a given facet.

to 71.2% as indexed by the $\Delta\chi^2$ criterion and from 44.6% to 48.8% as indexed by the ΔCFI criterion. The latter range is similar to the approximately 40% of DIF items detected in the original NEO-PI by Huang et al. (1997) using IRT methods.

In summary, the relatively small amount of DIF in the item loadings indicates that item-level metric equivalence was generally good for most facets. Thus, the behaviors referred to in most of the items are equally relevant indicators of the traits measured by the facet scales. In contrast, scalar equivalence (i.e., invariant item intercepts) was lacking for a substantial percentage of the items. This suggests that the behaviors referred to in the items are less prevalent, or are endorsed to a lesser extent, in some cultures than others for a given level of the associated trait.

Consistency of DIF across pairwise cultural comparisons. In the Appendix, we list—for each of the three pairwise cultural comparisons—the NEO-PI-R item numbers for those items exhibiting loading or intercept DIF based on the more conservative ΔCFI criterion. We also show which culture had the higher loading or intercept in each cultural comparison. There was a fair amount of overlap in the items exhibiting DIF across the three cultural comparisons. Of the total of 38 items that exhibited loading DIF on the basis of the ΔCFI criterion in one or more cultural comparisons, 13 (34.2%) exhibited loading DIF in more than one cultural comparison. Of the total of 170 items that exhibited intercept DIF on the basis of the ΔCFI criterion in one or more cultural comparisons, 99 (58.2%) exhibited intercept DIF in more than one cultural comparison. This indicates that some items are prone to DIF across multiple cultural comparisons, whereas other items exhibited DIF in only particular cultural comparisons.

We did not observe any definitive trend for DIF to be more or less frequent in any of the three pairwise cultural comparisons. However, we did observe that when loading DIF was detected,

Mexicans tended to have lower loadings in comparisons with Americans (see Table 3) and Filipinos (see Table 4), and Filipinos tended to have lower loadings than Americans (see Table 2). However, this was not always the case. Also, with the exceptions of two Openness to Experience facets (O4: Actions; O6: Values) and the A6: Tender-mindedness facet, we did not observe any consistent tendency for some facets to contain more DIF items or larger mean ΔCFI values than others across the three pairwise comparisons. These three facets were among those with the lowest alpha reliabilities in Mexico and the Philippines. Overall, DIF tended to be rather uniformly distributed across all domains and facets.

Interpretation of DIF. As previous researchers have noted, it is difficult to explain DIF items, for example, in terms of cultural norms, contexts, or practices (Ellis, 1990; Huang et al., 1997; Nye et al., 2008). Indeed, some interpretations may be speculative. We first examined whether the culture-specific item substitutions made in the Filipino and Spanish translations resulted in DIF. We focused on DIF based on the more conservative $\Delta CFI > .01$ criterion. Of the 10 item substitutions in the Filipino version, seven exhibited loading or intercept DIF in one or both comparisons involving the Filipino sample. The one culture-specific adaptation in the Spanish version (Las Vegas vs. Cancun) did not exhibit DIF in the United States-Mexican comparison, but did exhibit DIF with the Filipino item, which was also an item substitution referring to a willingness to try anything.

We next examined whether items whose keying had been reversed in the translations tended to exhibit DIF. Of 12 such items in the Filipino translation, all but one exhibited loading or intercept DIF in one or both cultural comparisons involving the Filipino sample. There were no such reversed items in the Spanish translation. In summary, most of the culture-specific item substitutions or reversals resulted in DIF in one or more cultural comparisons.

To determine whether some DIF might be due to translation inequivalencies, we had a Filipino–English bilingual and a Spanish–English bilingual carefully examine the Filipino and Mexican translations, respectively, and indicate any DIF items that might involve slight differences in meaning, as compared with the original English. On the basis of the bilinguals' judgments, a small number of additional cases of DIF may have been caused by translation inequivalencies, more so in comparisons involving the Filipino sample, in which direct or literal translation was apparently more difficult than in the Spanish translation. For example, in the E1: Warmth facet, Item 152 states: "I find it easy to smile and be outgoing with strangers." The Filipino translation, "*Madali sa 'king makisama sa mga di-kakilala*" ("It is easy for me to get along with strangers") is probably reasonable. However, the commonly used expression *makisama* (roughly meaning to get along) probably implies greater depth of interaction or relationship than is connoted by the English item, which could imply more superficial friendliness. This item exhibited intercept DIF, with Americans averaging higher than Filipinos.

Loading DIF. Aside from DIF that was attributed to item substitutions or translation inequivalencies, one additional pattern was observed for the loading DIF items. Items that involved negations (i.e., terms such as *not*, *don't*, and *rarely*) were more than twice as likely to exhibit loading DIF in one or more pairwise cultural comparisons, as compared with items that did not involve such negation terms. Almost one third of the inventory items that involved negations exhibited loading DIF, whereas only about 13% of the items without negations exhibited loading DIF. Thus, although the majority of items with negations did not exhibit loading DIF, the likelihood of loading DIF was greater with such items.

In some cases, the reasons for loading DIF may be more substantive or content-based. For example, a loading DIF item in the E5: Excitement-seeking facet makes reference to loving the excitement of roller coasters. The poor loading for this item in the Filipino sample indicates that this item is a poor indicator of excitement-seeking in the Philippine context, in which there are few roller coasters of any size. As another example, a loading DIF item in the A4: Compliance facet refers to one's hesitation to express anger even when justified. This item had a poor loading in the Philippines, as compared with the United States, perhaps because overt expression of anger is more strongly discouraged in the Philippines (Church, 1987; Lynch, 1973). Overall, however, aside from the items with negations, we were unable to identify consistent patterns or explanations for most of the small number of items that exhibited loading DIF.

Intercept DIF. We were able to discern plausible explanations for some intercept DIF. For example, the higher intercept in the Philippines, as compared with the United States, for an item in the N3: Depression facet about guilt or sinfulness might be explained by the greater religiosity (particularly Catholicism) of Filipinos. Filipinos also had higher intercepts than Americans on items that refer to pessimism about the future or to things looking hopeless. These differences might be due to the limited economic opportunities of many Filipinos, including college graduates. As another example, Mexicans had lower intercepts than Americans for two items in the E4: Activity facet that referred to being in a hurry or having a fast-paced life. These differences might be explained by the slower pace of life in Mexico (Levine & Norenzayan, 1999). These few examples illustrate that some intercept DIF is plausibly due to cultural differences in the prevalence or manifestation of specific attitudes and behaviors. However, like previous researchers, we found it difficult to discern definitive cultural explanations for many of the DIF items. Nonetheless, the substantial number of DIF items will make cross-cultural comparisons of facet scores risky, unless DIF cancels out at the facet-scale level.

Facet-level invariance: Does DIF cancel out or carry forward? As noted earlier, the question of whether DIF carries forward to the facet level of the NEO-PI-R is of central importance to personality psychologists, because it is at the facet level that aggregate scores are compared across cultures. Noninvariance at the facet level would call into question cross-cultural comparisons of the facet and Big Five domain scores.

We tested facet-level invariance by treating each Big Five dimension as a latent variable in separate multigroup CFA models with six facet scores as indicators of each Big Five trait. Noninvariant factor loadings would indicate that the given facet (e.g., N1: Anxiety) was not an equivalent or equally relevant indicator of the associated Big Five trait (e.g., Neuroticism) in two cultures. A noninvariant intercept would indicate that individuals with the same level of the Big Five trait, but from different cultural groups, averaged systematically lower or higher on the facet (Bollen, 1989). For example, if the intercept for the Self-consciousness facet of Neuroticism was higher in the Philippines than in the

United States, it would suggest that self-consciousness is a more prevalent manifestation of neuroticism in the Philippines than in the United States.

The nature and extent of facet-level invariance is summarized in the last four columns of Tables 2, 3, and 4. For each facet that exhibited noninvariance based on the chi-square difference test, we show the $\Delta\chi^2$ values (with one degree of freedom) obtained when the facet loading or intercept was constrained to be equal across cultures rather than freely estimated. For these facets, we also show the size of the CFI difference (ΔCFI) between the constrained and freely estimated models. Also indicated is the culture that had the higher facet loading or intercept when freely estimated. Note that loading invariance cannot be tested for facets that served as the anchor in the facet-level CFA models.

Loading noninvariance. As seen in Tables 2, 3, and 4, loading noninvariance was very infrequent at the facet level, especially using the ΔCFI $>$ .01 criterion. Across the three pairwise cultural comparisons, the percentage of facets that exhibited loading noninvariance ranged from 6.7% to 20.0% as indexed by the $\Delta\chi^2$ criterion and from 0% to 6.7% as indexed by the ΔCFI criterion. We might anticipate that loading noninvariance at the item level would translate to loading noninvariance at the facet level, because poor item indicators of the facet traits would detract from the quality of the facet scales as measures of the Big Five constructs. However, inspection of each facet in Tables 2, 3, and 4 reveals that loading DIF rarely carried forward to the facet level. This may be due to the relatively small size and amount of loading DIF. For example, in the United States-Philippines comparison in Table 2, three items in the C2: Order facet exhibited loading DIF on the basis of the $\Delta\chi^2$ criterion and two items on the basis of the ΔCFI criterion. However, the facet-level loadings for the C2: Order facet on the Big Five Conscientiousness construct were invariant across the two cultures.

Intercept noninvariance. In contrast to the infrequent loading noninvariance, intercept DIF frequently carried forward to the facet level. The percentages of facets that exhibited intercept noninvariance ranged from 33.3% to 73.3% as indexed by the $\Delta\chi^2$ criterion and from 20.0% to 56.7% as indexed by the ΔCFI criterion. Intercept DIF was particularly likely to carry forward when one of the two cultures had the higher item intercept for most of the DIF items and indices of model misfit (i.e., mean ΔCFI values) were substantial. For example, in Table 2, there were five items that exhibited intercept DIF (mean ΔCFI = .11) for the N4: Self-consciousness facet on the basis of the $\Delta\chi^2$ criterion, and the Filipino sample had a higher intercept than the American sample on four of those five items. This pattern of intercept DIF was then reflected in intercept noninvariance at the facet level ($\Delta\chi^2[1] = 33.68$; ΔCFI = .03), with the Filipino sample having a higher intercept than the Americans for the N4: Self-consciousness facet scale. Similarly, four items in the N5: Impulsivity facet exhibited intercept DIF (mean ΔCFI = .14), and the American sample averaged higher on three of the four items. This pattern of intercept DIF was then reflected in intercept noninvariance at the facet level ($\Delta\chi^2[1] = 23.70$; ΔCFI = .02), with Americans having a higher intercept than the Filipinos for the N5: Impulsivity facet scale.

In the comparisons of the American sample with the Filipino and Mexican samples, intercept DIF carried forward in this manner for a clear majority of the facets (see Tables 2 and 3). In the Philippine-Mexican comparison (see Table 4), intercept DIF carried forward to the facet level for a smaller percentage of the facets (about one third), probably because for many of the remaining facets, there was no consistent trend for one or the other culture to have the higher intercept for the DIF items in the facet. In these latter cases, intercept DIF can be viewed as having canceled out at the facet level (e.g., in Table 4, see facets E1, E4, O2, A1, A2, A3, C4, and C5). For three additional facets in the Philippine-Mexico comparisons (N4, N5, and E6), intercept DIF may appear to have canceled out at the facet level because no intercept noninvariance is shown for these facets in Table 4. However, recall that intercept invariance would not have been tested for these facets because they had already been shown to exhibit loading noninvariance.

In summary, the percentages of facets exhibiting some form of noninvariance (i.e., in either the loadings or intercepts) ranged from 53.3% to 80.0% as indexed by the $\Delta\chi^2$ criterion and from 26.7% to 56.7% as indexed by the ΔCFI criterion. These results indicate that a substantial amount of the DIF detected in the items, particularly in the item intercepts, was carried forward to the facet level.

Impact on cultural mean differences.

Original facet scales. Most of the original facet scales had statistically significant mean differences in the pairwise cultural comparisons. This was determined by conducting multivariate analyses of variance (MANOVAs) with culture as the independent variable and the 30 facet scales as dependent variables. Gender was introduced as a covariate in these analyses so that any significant effects would be attributable to culture rather than differences in the gender makeup of the three cultural samples (Johnson et al., 2008). The multivariate tests of cultural effects were statistically significant in each pairwise cultural comparison (Wilks's Λ range = .36–.55, $p < .01$). In follow-up ANOVAs of the individual facet scales, we observed significant ($p < .01$) cultural differences for 21 scales in the United States-Philippines comparison (partial η_p^2 range = .01–.32), 21 scales in the United States-Mexico comparison (η_p^2 range = .01–.08), and 16 scales in the Philippines-Mexico comparison (η_p^2 range = .01–.29).

Purified facet scales. The usual procedure when DIF is detected is to eliminate the DIF items before making cross-cultural comparisons (e.g., Huang et al., 1997; Johnson et al., 2008). To determine the impact of eliminating DIF items, we compared the size and significance of the cultural effects for the original facet scores and for "purified" facet scores derived by deleting DIF items with ΔCFI values greater than .01 (gender was again a covariate in these analyses). Different purified facet scales were derived for each pairwise cultural comparison, because the DIF items varied somewhat for each pairwise comparison. We retained only purified scales that had at least four items and alpha reliabilities of .50 or higher in both cultures being compared. Having done so, we found that a majority of the facet scales were no longer sufficiently reliable for cultural comparisons after eliminating DIF items. In all three pairwise cultural comparisons of the retained purified scales, the overall cultural effects were statistically significant (Wilks's Λ range = .89–.94, $p < .01$). Therefore, we compared the F statistics and effects sizes in follow-up ANOVAs with the original and purified scales.

In the comparison of the American and Filipino samples, purified scales of sufficient length and reliability were derived for 14 of the 30 facets. For six of these 14 facets, no cultural differences were observed for either the original or purified scales. Of the

remaining eight facets, all showed modest to moderate reductions in the size of the cultural effects with the purified scales as compared with the original scales (η_p^2 decreases of .01 to .06). For seven of these eight facets, the cultural differences that were statistically significant for the original scales were no longer significant for the purified scales. In summary, for eight of the 14 facets that could be compared, conclusions about the existence (i.e., statistical significance) or size of cultural mean differences would be different based on the purified scales, as compared with the original scales.

In the comparison of the American and Mexican samples, purified scales were derived for 13 facets (and one facet did not need to be purified). For six of these 13 facets, no cultural differences were observed for either the original or purified scales. Of the remaining seven facets, six showed modest to moderate reductions in the size of the cultural effects with the purified scales as compared with the original scales (η_p^2 decreases of .01 to .05). For three of these seven facets, the cultural differences that were statistically significant for the original scales were no longer significant for the purified scales. One facet scale exhibited significant cultural mean differences of comparable size for both the original and purified scales. In summary, for six of the 13 facets that could be compared, conclusions about the existence or size of cultural mean differences would be different based on the purified scales, as compared with the original scales.

Finally, in the comparison of the Filipino and Mexican samples, purified scales were derived for 11 facets. For six of these 11 facets, no cultural differences were observed for either the original or purified facet scales. Of the remaining five facets, three showed modest reductions in the size of the cultural effects (η_p^2 decreases of .01 to .03), although one of these effects was still statistically significant, and two showed slightly larger cultural effects for the purified scales as compared with the original scales. In summary, for five of the 11 facets that could be compared, conclusions about the size, direction, or existence of cultural mean differences would be different based on the purified scales, as compared with the original scales.

In summary, across the three pairwise cultural comparisons, conclusions about the statistical significance of cultural mean differences would have been different for 12 (31.6%) of the 38 possible comparisons of original and purified scales. For just under 50% of the comparisons, conclusions about the size of the cultural effects would have been different, although some of the effect size changes were modest. We could not discern any consistent patterns regarding which facets would have resulted in the same or different conclusions regarding cultural differences. Indeed, facets from all of the Big Five domains exhibited such changes between the original and purified scales in one or more of the cultural comparisons.

Taking into account invariance at the facet level. Although researchers typically remove DIF items from their scales before making cross-cultural comparisons, as described in the previous section, a case can be made for retaining the original scales in those cases in which DIF canceled out at the facet level. However, this is not a simple matter of retaining all of those facets that exhibited invariance in Tables 2, 3, and 4—that is, those facets for which DIF appeared to have cancelled out at the facet level. If other facet scales in the same Big Five domain were not invariant, then the facets defining the Big Five latent construct exhibited only partial invariance (Byrne et al., 1989). In particular, if many of the facets defining the same Big Five construct lack invariance, then it would call into question the cross-cultural equivalence of the latent Big Five domain scores used to test the invariance of the remaining facets, including those facets that appear to be invariant.

Researchers disagree on the proportion of invariant facets needed to enable cross-cultural comparisons. A rather liberal criterion was suggested by Byrne et al. (1989), who argue that as long as at least one indicator—in this case, one facet scale—beyond the anchor facet is invariant, sufficient partial invariance exists to permit cross-cultural comparisons. However, other researchers suggest that a partial invariance strategy should be used only when a large number of indicators (facets) are invariant (Brown, 2006; Vandenberg & Lance, 2000). In the present study, we used a moderately conservative strategy as follows: If in addition to the anchor facet (with the anchor intercept also being invariant), three of the five additional facets in the Big Five domain were invariant, then we considered it permissible to compare the four invariant facets across cultures, even in the presence of item-level DIF within these facets.

Using this partial invariance strategy, only a few additional facet scales would qualify for comparison based on the $\Delta\chi^2$ criterion—zero in the United States-Philippines comparison (see Table 2) and three in both the United States-Mexico and Philippine-Mexico comparisons (see Tables 3 and 4). In contrast, with the ΔCFI criterion, which detects less noninvariance, many more facet scales would qualify for comparison. In the United States-Philippines comparison, these include four Conscientiousness facets (C1, C2, C3, and C5; see Table 2). In the United States-Mexico comparison, these include five Neuroticism facets (N1, N2, N3, N4, and N6), five Extraversion facets (E2, E3, E4, E5, E6), and all six Conscientiousness facets (see Table 3). In the Philippines-Mexico comparison, these include four Neuroticism facets (N2, N4, N5, and N6) plus all of the Extraversion and Conscientiousness facets (see Table 4).

For several of these facet scales, the purified scales had been too short or unreliable to retain for cross-cultural comparisons. Thus, by using this partial invariance strategy, the number of facets that could be compared increased from 14 to 15 in the United States-Philippines comparison, from 14 to 20 in the United States-Mexico comparison, and from 11 to 19 for the Philippines-Mexico comparison. Thus, for each cultural comparison, there were still from 11 to 15 facets that could not be compared, either because they did not meet our criterion for partial invariance or because the purified scales were too short or unreliable.

Note that scales that qualified for comparison on the basis of our partial invariance strategy would no longer need to be purified of DIF items; DIF could be viewed as having cancelled out at the facet level. This would then reduce the number of scales showing different conclusions in tests of mean cultural differences, as compared with our earlier comparisons of original and purified scales.

Of the 15 facets that could now be compared in the United States-Philippines comparison, conclusions about the statistical significance and size of cultural differences would still change for seven and eight facets, respectively. Of the 20 facets that could now be compared in the United States-Mexico comparison, conclusions about the statistical significance and size of cultural differences would only change for one and three facets, respec-

tively. Of the 19 facets that could now be compared in the Philippines-Mexico comparison, conclusions about the statistical significance and size of cultural differences would only change for one and two facets, respectively. In summary, of 54 possible comparisons across the three pairwise cultural comparisons, conclusions about the significance and size of cultural mean differences would change for about 17% and 24% of the comparisons, respectively. These percentages are about half the size of those reported in our comparison of the original and purified scales. Keep in mind, however, that from one third to one half of the facet scales, depending on the particular pairwise cultural comparison, could still not be compared even in the partial invariance analysis.

Discussion

The primary goal of this study was to investigate the validity or meaningfulness of cross-cultural comparisons of mean personality profiles. Such comparisons have potential theoretical importance in increasing researchers' understanding of the ecological, cultural, and biological factors that influence personality. However, they have also generated controversy (e.g., Bock, 2000; Church, 2008; Poortinga et al., 2002). We focused on one important prerequisite for such comparisons, measurement invariance, using the NEO-PI-R, which is presently the most prominent inventory used in cross-cultural comparisons of personality profiles. Relatively few studies have investigated measurement invariance in personality inventories (e.g., Ellis et al., 1993; Ellis & Mead, 2000; Huang et al., 1997; Johnson et al., 2008; Waller et al., 2000), and only Huang et al. did so with a version of the NEO-PI. Thus, the present study is the first to examine measurement invariance in translated versions of the NEO-PI-R.

Our overall conclusion is that DIF is prevalent in the NEO-PI-R and frequently carries forward to the facet level. Using a partial invariance strategy can increase the number of facet scales that qualify for comparison across cultures, but still leaves many scales that cannot be confidently compared due to the presence of DIF or facet-level noninvariance. Thus, considerable caution is needed in drawing conclusions about mean trait differences between cultures on the basis of aggregate NEO-PI-R profiles.

Nature and extent of DIF. In total, about 40%–50% of the items in the NEO-PI-R exhibited some form of DIF across the three cultural comparisons. Approximately half of the DIF items exhibited DIF in more than one cultural comparison, revealing some replication of DIF and that some items are prone to DIF across multiple comparisons. One might argue that our primary criterion for identifying DIF (ΔCFI $>$ .01) is arbitrary and that different DIF criteria would result in different percentages of DIF items. It should be noted, however, that G. W. Cheung and Rensvold (2002) recommended this criterion as a way to identify differences in CFA model fit that have practical (not just statistical) significance. Indeed, as reported earlier, the DIF items identified using this criterion had cultural differences in freely estimated loadings and intercepts that were substantial in size. Thus, although the exact percentage of items showing DIF will, of course, depend on the significance level or other criterion applied, the items designated as DIF items in the present study exhibited nontrivial differences in loadings and intercepts.

The total percentage of DIF items in this study is slightly larger than the 40% reported by Huang et al. (1997) with the original NEO-PI and the approximately 30% reported by Johnson et al. (2008) with the MPQ. Johnson et al. examined DIF in a comparison of American and German samples, so it is possible that the greater proportion of DIF items in the present study was due, in part, to our inclusion of more diverse cultural samples. Careful back-translation procedures were used to derive the Filipino and Mexican versions of the NEO-PI-R, and our inspection of the DIF items suggested that translation inequivalence contributed to DIF for only a small number of items. In contrast, culture-specific item substitutions, although few in number, generally resulted in DIF. Although such adaptations may be necessary in some cases—and are not a problem for within-culture applications—our results suggest that they are likely to contribute to a lack of measurement invariance. Much of the remaining DIF is likely due to cultural differences in the relevance or prevalence (i.e., endorsement rate) of the behavioral indicators of the traits, although it can be difficult to explain these differences in terms of cultural norms, contexts, or practices (Ellis, 1990; Huang et al., 1997; Reise et al., 2001). The potential significance of this DIF for cultural comparisons is discussed in the following paragraphs.

Loading DIF. The most definitive finding of our study was that loading DIF was relatively infrequent and much less common than intercept DIF. This is important because it indicates that the thoughts, feelings, and behaviors referred to in the items are, in the vast majority of cases, equally relevant indicators of the associated traits in all three cultures. This provides strong evidence for the cross-cultural conceptual equivalence of the traits assessed by the facet scales and the Big Five dimensions. DIF in the factor loadings would represent a more fundamental problem for cross-cultural personality assessment, because it would indicate that the constructs being measured are defined differently, or manifested in different behaviors, in varied cultural contexts.

When loading DIF was detected, Mexicans tended to have lower loadings in comparisons with Americans and Filipinos, and Filipinos tended to have lower loadings than Americans, although this was not always the case. This is not a surprising finding, given that the items and associated behaviors were selected as relevant indicators of the traits for Americans. However, it does point to one benefit of DIF analyses, which can identify those behavioral exemplars of traits that are less relevant in assessing the traits in particular cultures.

Our finding that loading DIF was more likely for items that involved negations suggests that such items should be avoided in cross-cultural personality assessment. Test construction guidelines often call for inclusion of a balance of positively keyed and reverse-keyed items, in large part to offset the effects of acquiescence bias. However, if reverse-keyed items are obtained by using negations (e.g., "I don't find it easy to take charge"), rather than direct indicators of the opposite pole of the trait (e.g., "In meetings, I usually let others do the talking"), it may contribute to less invariant measurement across cultures.

Intercept DIF. In contrast to loading DIF, intercept DIF was fairly common—about 40% of the items based on the more conservative ΔCFI criterion. This indicates that there were cultural differences in the extent to which the thoughts, feelings, and behaviors referred to in these items were endorsed by respondents. It should be noted that a lack of intercept invariance could reflect either systematic measurement bias or valid group differences in trait levels (Cole, Maxwell, Arvey, & Salas, 1993; Hancock,

1997). Systematic bias would indicate that individuals with the same level of the latent trait but from different cultures rate the item differently, leading to noninvariant intercepts. However, it is also possible that respondents in a particular culture endorse an item less, and hence have lower intercepts, because they actually average lower on the latent trait. This latter possibility is more likely when the same culture shows lower intercepts on many of the items in the facet scale, revealing a fairly uniform pattern of lower endorsement of the behavioral exemplars of the trait (i.e., analogous to a cultural main effect). In contrast, when there is no consistent or uniform pattern of DIF within a facet, DIF more likely reflects bias associated with the content of specific DIF items (i.e., analogous to a Culture × Item interaction).

For many of the facet scales in our pairwise cultural comparisons, a fairly uniform pattern of DIF was exhibited, suggesting that some DIF was due to average latent trait differences. For many other facets, however, this was not the case. That is, there was no consistent trend for one or the other culture to have the higher intercept for the DIF items in the facet. Furthermore, many apparent cultural mean differences were reduced or eliminated when DIF items were removed. These findings suggest that at least some DIF was, in fact, due to measurement bias, that is, cultural differences in the relevance or prevalence of the specific behaviors (items) used as indicators of the respective traits. Unfortunately, elimination of DIF items to remove such biases can also raise questions about the content-representativeness of the remaining items and thus descriptions of manifest trait levels in each culture. Certainly, more confident conclusions about cultural differences would be possible if relatively few items exhibit DIF, which was not the case here.

Does DIF carry forward to the facet scale level? As noted earlier, McCrae, Terracciano, & 79 Members (2005b) suggested that DIF might cancel out at the facet level, allowing valid cross-cultural comparisons even in the presence of DIF. A second goal of our study was to investigate this possibility. A few studies with real and simulated data have suggested that DIF may not cancel out at the scale level in cross-cultural comparisons (e.g., Ellis & Mead, 2000; Li & Zumbo, 2009; Nye et al., 2008; see, however, Waller et al., 2000). However, none of these studies investigated this question with the NEO-PI-R. Our results were clear. McCrae, Terracciano, & 79 Member's (2005b) speculation was generally supported in the case of loading DIF, but not intercept DIF.

Indeed, loading DIF rarely contributed to loading noninvariance at the facet level. Of course, loading DIF was small and infrequent, so it is possible that loading DIF will carry forward to the scale level when loading DIF is more extensive. Furthermore, loading DIF can also contribute to intercept noninvariance at the facet level.[3] In contrast, intercept DIF frequently carried forward, contributing to intercept noninvariance at the facet level. Our partial invariance analysis suggested that some intercept DIF cancelled out at the facet level, making more cross-cultural comparisons possible. However, this was not the case for many other facets, and many purified scales were too short or unreliable for cross-cultural comparisons. Furthermore, our finding that some cultural differences were reduced or eliminated after deleting DIF items suggests that item bias was carried forward for at least some facets. At a minimum, our results imply that considerable caution is necessary in interpreting profile differences across cultures, particularly when the differences are not large.

How might we reconcile our findings with the results of studies that suggest such profile comparisons may be meaningful? As noted earlier, substantial (though imperfect) geographical patterning of personality profiles, and sensible (although not consistently replicable) country-level correlates of Big Five scores, have been observed (e.g., Allik & McCrae, 2004; Hofstede & McCrae, 2004; McCrae & Terracciano, 2008; McCrae et al., 2010; Schmitt et al., 2007). In addition, country-level means are generalizable across gender and age groups in both self-report and observer data (Costa et al., 2001; McCrae, 2001, 2002; McCrae, Terracciano, & 78 Members, 2005a; McCrae et al., 2010; Schmitt et al., 2007). Several explanations seem plausible.

One possibility, examined in this study, is that intercept noninvariance cancels out at the facet level or is caused by valid cultural differences in trait levels. Our results suggest that this may be the case for some but not all facets. A second possibility is that cultural differences in trait levels are sufficiently large to overshadow any biases associated with DIF. Although only 4% of the variance in NEO-PI-R scores is accounted for by culture (McCrae & Terracciano, 2008), the scores for some facets do show nontrivial departures from American norms in some cultures (see, e.g., McCrae, 2002, Appendix I). A third possibility is that various sources of bias (e.g., DIF, response styles) that impact personality scores across cultures also impact measures of other constructs with which these personality scores are correlated. Indeed, most of the constructs that have been related to country-level NEO-PI-R scores have also been measured using self-report, so they would be subject to similar response biases (e.g., Gelade et al., 2006; Hofstede & McCrae, 2004; Leung & Bond, 2004; Schmitt et al., 2007). Finally, some culture-level studies have focused on the Big Five domain scores rather than the facet scores (Hofstede & McCrae, 2004; Schmitt et al., 2007). It is possible that facet-level noninvariance cancels out at the Big Five domain level, although this seems unlikely given our finding that at least some DIF carried forward to the facet level. Also, McCrae et al. (2010) recently argued that cross-cultural comparisons of aggregate traits should be conducted at the facet level rather than the domain level.

Each of these explanations may have some validity and additional research will be needed to determine their relative efficacy. One recent approach—which avoids some of the problems associated with shared method variance—has been to correlate NEO-PI-R profile scores with nation-level indices of relevant behaviors (e.g., pace of life, longevity, suicide rates, alcohol consumption, and corruption; see, e.g., Heine, Buchtel, & Norenazyan, 2008; McCrae et al., 2010; Mõttus, Allik, & Realo, 2010; Oishi & Roth, 2009). For example, Heine et al. (2008) found that national character ratings (i.e., ratings of typical personality) predicted country-level indices of Conscientiousness better than did cultural means for self-reported and observer-reported Conscientiousness, suggesting that cultural mean profiles may be invalid. Oishi and Roth

[3] Loading DIF could impact intercept invariance at the facet level because items with different item factor loadings (i.e., regression slopes) in two cultures will tend to have different item intercepts as well (although intercept DIF is not formally tested when loading DIF is present). However, with the possible exception of the O4: Actions facet in the United States-Mexican comparison (see Table 3), loading DIF did not appear to account for intercept noninvariance at the facet level.

(2009) replicated Heine et al.'s findings for Conscientiousness but did find that country mean scores for Agreeableness and Neuroticism predicted relevant country-level behavioral indices. However, Mõttus et al. noted a number of complexities in conducting such studies. For example, because different facets of Conscientiousness (and probably other Big Five traits) apparently relate very differently to culture-level criteria, it will be important to relate aggregate profiles and culture-level criteria at the facet level rather than the domain level. In addition, these relationships apparently depend on whether the aggregate facet scores are based on self- or observer ratings (Mõttus et al., 2010). Noting such complexities, and the often theoretically loose hypotheses linking personality profiles to culture-level criteria, Mõttus et al. concluded that "previous research on the predictive validity of nation-level mean personality scores has not been theoretically and methodologically rigorous enough to warrant *definitive* conclusions" (p. 639). In summary, it is too early to tell whether such studies will be able to resolve the apparent discrepancy between DIF results and other evidence suggesting that aggregate personality profiles may be meaningful.

Practical implications. What are some practical implications of our findings? On the basis of our results, and those of previous studies, it is apparent that a substantial proportion of items in personality inventories—typically about 30%–50%—will exhibit DIF, and some DIF will carry forward to the facet or score level. Although not a desirable amount of DIF, this percentage range might be viewed as somewhat normative for carefully back-translated inventories and a target to improve on in subsequent translation efforts. Given that DIF was rather uniformly distributed across all domains and facets, eliminating DIF will not be a simple matter of revising the items in specific facet scales. Extensive DIF may not be a problem for within-culture applications of personality inventories, although it will be best to develop local norms for applied use, rather than relying on American norms. In addition, if the inventory will be used for mean comparisons of cultural or ethnic subgroups within a culture, DIF analyses should be extended to such subgroup comparisons. Cross-cultural or cross-ethnic comparisons of nomological networks (e.g., behavioral correlates) and mean trait levels may be misleading in the presence of substantial DIF. In addition to DIF analyses, we recommend that facet-level invariance also be examined. Using a partial invariance strategy, it may be possible to justify cross-cultural comparisons with additional scales. Unfortunately, there is presently little consensus regarding the number of invariant facets needed to implement such a strategy, and more research is needed in this regard (Brown, 2006; Byrne et al., 1989; Vandenberg & Lance, 2000).

In theory, one way that researchers might attempt to avoid the problems associated with noninvariant measures would be to apply a culturally decentered approach. Researchers could try to construct inventories simultaneously in several cultures and select items that have invariant loadings and intercepts. However, Strelau and Angleitner's (1994) experience in developing the Pavlovian Temperament Survey may reveal the challenges associated with this approach. The researchers developed a pool of 252 items, from which researchers selected subsets of items with good item discrimination in their respective countries. It is revealing, however, that only about 50% of the selected items overlapped in any two cultural versions of the instrument. These results indicate that it will be difficult to construct culturally decentered measures that contain only universal or invariant indicators of traits.

Presently, if cross-cultural comparisons are desired, one recommendation would be to use scales that have many items, so that DIF items can be eliminated without substantial reductions in reliability and content representativeness. It is premature, however, to recommend that DIF items in existing inventories always be eliminated before making cross-cultural comparisons. One reason is that elimination of DIF items could result in scales that are too short, unreliable, or lacking in content representativeness. A second reason is that more studies are needed to reconcile the typical finding of substantial DIF with culture-level findings, which suggest that cross-cultural mean comparisons may be meaningful (e.g., Allik & McCrae, 2004; Hofstede & McCrae, 2004). Indeed, DIF analyses alone cannot tell us whether scale scores are valid or accurate, but they do tell us whether the scales provide equivalent measurement across cultures. Presently, definitive conclusions about the validity of cross-cultural mean comparisons are not possible, although our findings should caution researchers against taking too strong a stance in favor of their validity. We recommend that researchers use—and attempt to reconcile—both "bottom-up" approaches (e.g., DIF studies) and "top-down" approaches (e.g., studies of culture-level correlates of mean profiles) in cross-cultural studies. Hopefully, this combined approach will eventually lead to more confident conclusions regarding the validity of mean profile comparisons with particular instruments.

Strengths and limitations. Strengths of the study included the diverse cultural samples investigated, the examination of invariance at both item and facet levels, and the fact that DIF was investigated in three pairwise cultural comparisons. There were also some limitations. First, our total sample sizes were sufficient for CFA analyses, but were not large enough to investigate DIF separately for men and women (which in any case is not typically done) or to match the cultural samples on the proportion of men and women (e.g., Johnson et al., 2008). We did control for gender differences in the MANOVAs that tested for cultural mean differences in the original and purified scales.

Second, some authors have recommended that researchers apply iterative DIF procedures in which items initially identified as DIF are eliminated and the remaining items are tested again for DIF (e.g., Huang et al., 1997; Park & Lautenschlager, 1990). We did not use an iterative process, in part, because there would not have been a large number of items remaining in each facet scale after deletion of the initial DIF items. In addition, our goal was not to develop a more refined measure but rather to draw conclusions about the extent of DIF in the NEO-PI-R. Although iterative procedures might change the exact number of DIF items, it is unlikely that our overall conclusions would have changed.

Third, although it is commonplace for researchers to consider cultural explanations of DIF, this may be premature without first replicating the DIF items in additional samples. Also, other factors besides language and culture (e.g., differential motivation of samples, acquiescence bias) could contribute to DIF. Of course, such factors could also contribute to noninvariance in the samples now being compared in multinational studies of personality profiles.

A final limitation was our focus on an imported or "imposed-etic" measure (Berry, 1969; Church, 2001). Imported measures such as the NEO-PI-R may miss salient culture-unique behavioral exemplars of the traits being assessed. Only indigenous instru-

ments, or measures developed using a combined "emic-etic" (indigenous-imported) approach can identify and incorporate such behaviors (e.g., F. M. Cheung, Cheung, Wada, & Zhang, 2003; Katigbak et al., 2002; Ortiz et al., 2007). Of course, an important drawback of indigenous approaches is that truly culture-unique items, by their very nature, are unlikely to exhibit measurement invariance across cultures.

Conclusion

Although cross-cultural comparisons of personality profiles have the potential to clarify the ecological, cultural, and biological bases of personality, the validity of such comparisons is still unresolved in our view. Although a number of factors can confound such comparisons (e.g., sampling differences, reference group effects), the present study focused on one important prerequisite—measurement invariance. DIF studies of personality measures are still rare. However, the available evidence indicates that there is substantial DIF in major personality inventories and that cultural differences are typically reduced or eliminated after DIF items are removed (Huang et al., 1997; Johnson et al., 2008; Nye et al., 2008). Further research on the impact of DIF on score comparisons is needed, using both real and simulated data. One goal of such research would be to reconcile the presence of extensive DIF—which does not always cancel out at the scale level—with other findings (e.g., geographical patterning, external correlates), which suggest that cultural mean profiles are reasonably valid or accurate. In the meantime, the take-home message of the present study is that considerable caution is warranted in cross-cultural comparison of personality profiles.

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. doi:10.1177/0022022103260382

Benet-Martinez, V., & John, O. P. (1998). Los cinco grandes across culture and ethnic groups: Multitrait multimethod analysis of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75,* 729–750.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246. doi:10.1037/0033-2909.107.2.238

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4,* 119–128. doi:10.1080/00207596908247261

Bock, P. K. (2000). Culture and personality revisited. *American Behavioral Scientist, 44,* 32–40. doi:10.1177/00027640021956071

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: John Wiley.

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

Butcher, J. N. (Ed.). (1996). *International adaptations of the MMPI-2.* Minneapolis: University of Minnesota Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466. doi:10.1037/0033-2909.105.3.456

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

Camperio Ciani, A. S., Capiluppi, C., Veronese, A., & Sartori, G. (2007). The adaptive value of personality differences revealed by small island population dynamics. *European Journal of Personality, 21,* 3–22. doi: 10.1002/per.595

Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (1993). *Sixteen Personality Factor Questionnaire* (5th ed.). Champaign, IL: Institute for Personality and Ability Testing.

Cheung, F. M., Cheung, S. F., Wada, S., & Zhang, J. (2003). Indigenous measures of personality assessment in Asian countries: A review. *Psychological Assessment, 15,* 280–289. doi:10.1037/1040-3590.15.3.280

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9,* 233–255. doi:10.1207/S15328007SEM0902_5

Church, A. T. (1987). Personality research in a non-Western culture: The Philippines. *Psychological Bulletin, 102,* 272–292. doi:10.1037/0033-2909.102.2.272

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69,* 979–1006. doi:10.1111/1467-6494.696172

Church, A. T. (2008). Current controversies in the study of personality across cultures. *Social and Personality Psychology Compass, 2,* 1930–1951. doi:10.1111/j.1751-9004.2008.00132.x

Church, A. T. (2010). Measurement issues in cross-cultural research. In G. Walford, M. Viswanathan, & E. Tucker (Eds.), *The Sage handbook of measurement* (pp. 151–177). Thousand Oaks, CA: Sage Publications.

Church, A. T., & Katigbak, M. S. (2002). The five-factor model in the Philippines: Investigating trait structure and levels across cultures. In R. R. McCrae & J. Allik (Eds.), *The five-factor model across cultures* (pp. 129–154). New York, NY: Kluwer Academic/Plenum Publishers.

Church, A. T., Katigbak, M. S., Miramontes, L. G., del Prado, A. M., & Cabrera, H. F. (2007). Culture and the behavioral manifestations of traits: An application of the act frequency approach. *European Journal of Personality, 21,* 389–417. doi:10.1002/per.631

Church, A. T., Katigbak, M. S., Reyes, J. A. S., Salanga, M. G. C., Miramontes, L. A., & Adams, N. B. (2008). Prediction and cross-situational consistency of daily behavior across cultures: Testing trait and cultural psychology perspectives. *Journal of Research in Personality, 42,* 1199–1215. doi:10.1016/j.jrp.2008.03.007

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114,* 174–184. doi:10.1037/0033-2909.114.1.174

Costa, P. T. Jr., & 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.

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

Del Pilar, G. H. (1998). *L'extraversion en psychologie differentielle et l'extratension au Test De Rorschach* [Extraversion in differential psychology and Rorschach extratension] (Doctoral dissertation, University of Paris X-Nanterre).

Ellis, B. B. (1990). Extension and evaluation of the Humphreys/Hulin emic item identification hypothesis using item response theory. In P. J. D. Drenth, J. A. Sergeant, & R. J. Takens (Eds.), *European perspectives in psychology* (Vol. 1, pp. 43–55). New York, NY: John Wiley.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology, 24,* 133–148. doi:10.1177/0022022193242001

Ellis, B. B., & Mead, A. D. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF Questionnaire. *Educational and Psychological Measurement, 60,* 787–807. doi:10.1177/00131640021970781

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67,* 565–582. doi:10.1177/0013164406296975

Finch, W. H., & French, B. F. (2008). Using exploratory factor analysis for locating invariant referents in factor invariance studies. *Journal of Modern Applied Statistical Methods, 7,* 223–233.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 13,* 378–402. doi:10.1207/s15328007sem1303_3

French, B. F., & Finch, W. H. (2008, March). *When under the influence of non-invariant factor loadings, does computation method of the factor score matter?* Paper presented at the American Educational Research Association, New York, NY.

French, B. F., & Maller, S. J. (2006, April). *The influence of differential item functioning on internal consistency reliability.* Paper presented at the American Educational Research Association, San Francisco, CA.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for DIF detection. *Educational and Psychological Measurement, 67,* 373–393. doi:10.1177/0013164406294781

French, B. F., Maller, S. J., & Zumbo, B. (2007, April). *The influence of differential item functioning on multi-sample confirmatory factor analysis.* Paper presented at the National Council on Measurement in Education conference, Chicago, IL.

Gelade, G. A., Dobson, P., & Gilbert, P. (2006). National differences in organizational commitment: Effect of economy, product of personality, or consequence of culture? *Journal of Cross-Cultural Psychology, 37,* 542–556. doi:10.1177/0022022106290477

Gellman, M. (1994). *The revised NEO Personality Inventory: Manual supplement for the Spanish edition.* Odessa, FL: Psychological Assessment Resources.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development, 30,* 91–105.

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. doi:10.1111/j.1467-9280.2008.02085.x

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. doi:10.1037/0022-3514.82.6.903

Hofstede, G., & McCrae, R. R. (2004). Personality and culture revisited: Linking traits and dimensions of culture. *Cross-Cultural Research, 38,* 52–88. doi:10.1177/1069397103259443

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology, 28,* 192–218. doi:10.1177/0022022197282004

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. doi:10.1111/j.1467-6494.2008.00500.x

Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice: Principles, advances, and applications. *The Counseling Psychologist, 34,* 684–718. doi:10.1177/0011000006286347

Katigbak, M. S., Church, A. T., Guanzon-Lapeña, M. A., Carlota, A. J., & del Pilar, G. H. (2002). Are indigenous personality dimensions culture specific? Philippine inventories and the five-factor model. *Journal of Personality and Social Psychology, 82,* 89–101. doi:10.1037/0022-3514.82.1.89

Labouvie, E., & Ruetsch, C. (1995). Testing for equivalence of measurement scales: Simple structure and metric invariance reconsidered. *Multivariate Behavioral Research, 30,* 63–76. doi:10.1207/s15327906mbr3001_4

Leung, K., & Bond, M. H. (2004). Social axioms: A model for social beliefs in multicultural perspective. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 36, pp. 119–197). San Diego, CA: Elsevier Academic Press.

Levine, R. V., & Norenzayan, A. (1999). The pace of life in 31 countries. *Journal of Cross-Cultural Psychology, 30,* 178–205. doi:10.1177/0022022199030002003

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica, 30,* 343–370.

Lynch, R. (1973). Social acceptance reconsidered. In F. Lynch & A. de Guzman II (Eds.), *Four readings on Philippine values* (4th ed., enlarged, Institute of Philippine Culture Papers No. 2, pp. 1–68). Quezon City, Philippines: Ateneo de Manila University Press.

McCrae, R. R. (2001). Trait psychology and culture: Exploring intercultural comparisons. *Journal of Personality, 69,* 819–846. doi:10.1111/1467-6494.696166

McCrae, R. R. (2002). NEO-PI-R data from 36 cultures: Further intercultural comparisons. In R. R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 105–125). New York, NY: Kluwer Academic/Plenum Publishers.

McCrae, R. R. (2004). Human nature and culture: A trait perspective. *Journal of Research in Personality, 38,* 3–14. doi:10.1016/j.jrp.2003.09.009

McCrae, R. R., & Allik, J. (Eds.). (2002). *The five-factor model across cultures.* New York, NY: Kluwer/Plenum.

McCrae, R. R., Costa, P. T., Jr., del Pilar, G. Y., Rolland, J.-P., & Parker, W. D. (1998). Cross-cultural assessment of the five-factor model: The Revised NEO Personality Inventory. *Journal of Cross-Cultural Psychology, 29,* 171–188. doi:10.1177/0022022198291009

McCrae, R. R., & Terracciano, A. (2008). The five-factor model and its correlates in individuals and cultures. In F. J. R. van de Vijver, D. A. van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analyses of individuals and cultures* (pp. 249–283). Mahwah, NJ: Erlbaum.

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

McCrae, R. R., Terracciano, A., & 79 Members of the Personality Profiles of Cultures Project. (2005b). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89,* 407–425. doi:10.1037/0022-3514.89.3.407

McCrae, R. R., Terracciano, A., De Fruyt, F., De Bolle, M., Gelfand, M. J., Costa, P. T., Jr., & 42 Collaborators of the Adolescent Personality Profiles of Cultures Project. (2010). The validity and structure of culture-level personality scores: Data from ratings of young adolescents. *Journal of Personality, 78,* 815–838. doi:10.1111/j.1467-6494.2010.00634.x

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7,* 361–388. doi:10.1177/1094428104268027

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. doi:10.1016/j.jrp.2010.08.005

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York, NY: McGraw-Hill.

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

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. doi:10.1016/j.jrp.2008.11.002

Olson, K. R. (2007). Why do geographic differences exist in the worldwide distribution of extraversion and openness to experience? The history of human emigration as an explanation. *Individual Differences Research, 5,* 275–288.

Ortiz, F. A., Church, A. T., Vargas-Flores, J., Ibanez-Reyes, J., Flores-Galaz, M., Iuit-Briceño, J. I., & Escamilla, J. M. (2007). Are indigenous personality dimensions culture-specific? Mexican inventories and the five-factor model. *Journal of Research in Personality, 41,* 618–649. doi:10.1016/j.jrp.2006.07.002

Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14,* 163–173. doi:10.1177/014662169001400205

Poortinga, Y. H., Van de Vijver, F. J. R., & Van Hemert, D. A. (2002). Cross-cultural equivalence of the Big Five: A tentative interpretation of the evidence. In R. R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 281–302). New York, NY: Kluwer Academic/Plenum Publishers.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529. doi:10.1037/0021-9010.87.3.517

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI-R Neuroticism scale. *Multivariate Behavioral Research, 36,* 83–110. doi:10.1207/S15327906MBR3601_04

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

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

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, 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. doi:10.1177/0022022106297299

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91,* 1292–1306. doi:10.1037/0021-9010.91.6.1292

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78–107. doi:10.1086/209528

Strelau, J., & Angleitner, A. (1994). Cross-cultural studies on temperament: Theoretical considerations and empirical studies based on the Pavlovian Temperament Survey. *Personality and Individual Differences, 16,* 331–342. doi:10.1016/0191-8869(94)90170-8

Taylor, T. R., & Boeyens, J. C. (1991). The comparability of scores of Blacks and Whites on the South African Personality Questionnaire: An exploratory study. *South African Journal of Psychology, 21,* 1–11.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *Handbook of personality theory and testing: Personality measurement and assessment* (Vol. 2, pp. 261–292). London, England: Sage.

Terracciano, A., Abdel-Khalek, A. M., Ádám, N., Adamovová, L., Ahn, C.-K., Ahn, H.-N., . . . McCrae, R. R. (2005, October 7). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310,* 96–100. doi:10.1126/science.1117199

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. doi:10.1177/109442810031002

van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

van Leest, P. F. (1997). Bias and equivalence research in the Netherlands. *European Review of Applied Psychology/Revue Européenne de Psychologie Appliquée, 47,* 319–329.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5,* 125–146. doi:10.1037/1082-989X.5.1.125

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analysis? Implications for translating language tests. *Language Testing, 20,* 136–147. doi:10.1191/0265532203lt248oa

Zumbo, B. D., & Koh, K. H. (2005). Manifestation of differences in item-level characteristics in scale-level measurement invariance tests of multi-group confirmatory factor analysis. *Journal of Modern Applied Statistical Methods, 4,* 275–282.

(*Appendix follows*)

Appendix

NEO-PI-R Items With Loading and Intercept DIF in One or More Cultural Comparisons

	Loading DIF			Intercept DIF		
NEO-PI-R facet	US-Phil	US-Mex	Phil-Mex	US-Phil	US-Mex	Phil-Mex
N1: Anxiety						
1	—	US>M	P>M	—	—	—
31	—	—	—	—	US>M	P>M
61	—	US>M	—	—	—	P>M
91	—	—	—	—	US<M	—
121	—	—	—	US<P	—	P>M
151	—	US>M	P>M	—	—	—
181	—	—	—	US>P	—	—
211	—	—	—	—	US>M	—
N2: Angry Hostility						
36	—	—	—	—	—	P<M
66	—	—	—	US<P	US<M	—
96	—	—	P>M	US<P	—	—
156	—	—	—	US<P	US<M	—
186	—	US<M	—	—	—	—
216	—	—	—	US>P	—	—
N3: Depression						
11	—	US>M	P>M	—	—	—
101	—	—	—	US<P	US>M	P>M
131	—	—	—	—	US>M	P>M
161	—	—	—	—	US>M	P>M
191	—	—	—	US<P	—	—
N4: Self-Consciousness						
16	—	—	—	US<P	—	P>M
46	—	—	—	—	US>M	P>M
76	—	—	—	US<P	—	P>M
106	—	—	—	US>P	US<M	P<M
136	—	—	—	—	US>M	P>M
166	—	—	—		US<M	—
196	—	—	—	US<P	US>M	P>M
226	—	—	—	—	—	P>M
N5: Impulsiveness						
21	—	US>M	—	US>P	—	P>M
51	—	—	—	—	US>M	P>M
111	—	—	—	US<P	—	P>M
141	—	—	P>M	US>P	US>M	—
171	—	—	—	—	US<M	P<M
231	—	—	—	US>P	—	P<M
N6: Vulnerability						
56	—	—	—	US<P	—	—
146	—	—	—	—	US>M	P>M
176	—	—	—	—	US<M	P<M
206	—	—	—	—	—	P>M
236	—	—	—	US<P	US<M	—
E1: Warmth						
2	—	—	—	US>P	US>M	P>M
32	—	—	—	US>P	US>M	—
62	—	—	—	US>P	US>M	—
92	—	—	—	US>P	US>M	P<M
152	—	—	—	US>P	—	—
212	—	—	—	—	US>M	P>M
E2: Gregariousness						
7	—	—	—	—	US<M	—
37	—	—	—	—	US<M	—
97	—	—	—	—	US>M	P>M
127	—	—	—	US>P	US>M	—

(*Appendix continues*)

Appendix (*continued*)

	Loading DIF			Intercept DIF		
NEO-PI-R facet	US-Phil	US-Mex	Phil-Mex	US-Phil	US-Mex	Phil-Mex
E3: Assertiveness						
12	—	—	—	—	US<M	—
42	—	—	—	US<P	US<M	—
72	—	—	—	US>P	US>M	—
162	US>P	—	—	—	US>M	P<M
192	—	—	—	—	—	P<M
222	US>P	—	—	—	—	—
E4: Activity						
47	—	—	—	—	US<M	—
77	—	—	—	US>P	US>M	—
107	—	—	—	—	US>M	P>M
137	—	—	—	US>P	—	P<M
167	—	—	—	—	US>M	—
197	—	—	—	—	US>M	P>M
227	—	—	—	US>P	—	P<M
E5: Excitement-Seeking						
22	—	—	—	US>P	US>M	P<M
52	—	—	—	—	—	P<M
82	—	—	P<M	—	—	—
112	—	—	—	—	US<M	P<M
142	—	—	P<M	US>P	—	—
172	—	—	P<M	US>P	—	—
202	—	—	—	—	—	P>M
232	—	—	—	US>P	US>M	P>M
E6: Positive Emotions						
87	—	—	—	US>P	—	P<M
207	—	—	—	—	—	P>M
237	—	—	—	US>P	—	—
O1: Fantasy						
3	—	—	—	—	—	P<M
33	—	—	—	US>P	US>M	—
63	—	—	—	US<P	—	—
93	—	—	—	US>P	—	—
123	—	—	—	—	US<M	—
153	—	—	P>M	—	US>M	—
183	—	—	—	US>P	—	—
213	—	—	—	—	US>M	—
O2: Aesthetics						
8	—	—	—	US<P	US<M	—
38	—	—	—	US>P	US>M	—
68	—	—	—	US<P	US<M	—
98	—	—	—	US<P	—	P>M
158	—	—	—	US<P	—	P>M
218	—	—	—	US<P	US<M	—
O3: Feelings						
13	—	—	—	—	US>M	P>M
43	—	—	—	US>P	US>M	—
73	—	—	—	US>P	US>M	—
163	—	—	—	US<P	—	P>M
193	—	—	—	—	US>M	P>M
223	—	—	—	US<P	US>M	P>M
O4: Actions						
18	—	US>M	P>M	US<P	—	—
48	—	US<M	—	—	—	P<M
78	—	US>M	P>M	US<P	—	—
108	—	—	—	US<P	—	P>M
138	—	US>M	P>M	—	—	—
168	—	—	—	—	—	P<M
198	—	—	—	US>P	US<M	P<M
228	—	US>M	—	—	—	—

(*Appendix continues*)

Appendix (*continued*)

NEO-PI-R facet	Loading DIF			Intercept DIF		
	US-Phil	US-Mex	Phil-Mex	US-Phil	US-Mex	Phil-Mex
O5: Ideas						
113	—	—	—	—	US<M	—
143	—	—	—	US<P	—	—
O6: Values						
28	—	—	—	US>P	US>M	P<M
58	—	—	—	US>P	US<M	P<M
88	—	US>M	—	US>P	—	P<M
118	—	—	P<M	US>P	US<M	—
148	—	US>M	P>M	—	—	—
178	—	—	—	US>P	—	P<M
208	—	US>M	P>M	US>P	—	—
238	—	—	—	US>P	—	P<M
A1: Trust						
4	—	—	—	—	—	P<M
34	—	—	—	—	US>M	—
64	US>P	—	—	—	US>M	—
124	—	—	—	US>P	—	—
154	—	—	—	—	US>M	—
184	—	—	—	—	US>M	—
214	—	—	—	—	—	P>M
A2: Straightforwardness						
9	—	—	—	US<P	US>M	—
39	—	—	—	US>P	—	P<M
129	—	—	—	US>P	US>M	—
159	US>P	—	P<M	—	—	—
189	—	—	—	US>P	—	P<M
219	—	—	—	US>P	US>M	—
A3: Altruism						
14	—	—	—	US>P	—	P<M
44	—	—	—	—	US>M	—
74	—	—	—	US>P	US>M	P<M
134	—	—	—	—	US>M	—
164	—	—	—	US>P	US>M	—
194	US<P	—	—	—	—	P>M
224	—	—	—	—	—	P>M
A4: Compliance						
19	—	—	—	US<P	—	P>M
49	US<P	—	—	—	—	P>M
79	US>P	—	—	—	—	—
109	—	—	—	US>P	—	—
139	—	—	—	US<P	US>M	P>M
169	—	—	—	—	US>M	P>M
199	—	—	—	US<P	—	—
229	—	—	—	US>P	—	—
A5: Modesty						
24	—	—	—	US>P	US>M	—
54	—	—	—	—	US>M	P>M
84	—	—	—	US>P	US>M	—
114	—	US>M	—	—	—	—
144	—	—	—	—	US>M	P>M
174	—	—	—	—	US>M	P>M
204	—	—	—	—	US>M	P>M
234	—	—	—	US>P	—	P<M
A6: Tender-Mindedness						
29	—	—	—	US>P	US<M	P<M
59	—	—	—	US<P	US<M	—
89	—	US>M	P>M	—	—	—
119	—	—	—	US<P	US<M	P>M
149	US>P	—	—	—	US<M	—
179	—	—	—	US>P	US<M	P<M
209	—	—	—	—	US>M	P>M
239	—	—	—	US<P	US>M	—

(*Appendix continues*)

Appendix (*continued*)

	Loading DIF			Intercept DIF		
NEO-PI-R facet	US-Phil	US-Mex	Phil-Mex	US-Phil	US-Mex	Phil-Mex
C1: Competence						
5	—	—	—	US<P	—	P>M
65	—	—	—	US>P	—	—
125	—	—	—	US>P	—	P<M
185	—	—	—	US>P	—	P<M
215	—	—	—	US>P	—	P<M
C2: Order						
10	—	—	P>M	—	US>M	—
70	—	—	P>M	—	—	—
100	—	—	P>M	—	—	—
130	—	—	—	—	—	P>M
160	US<P	—	P>M	—	—	—
190	US>P	US>M	—	—	—	P>M
220	—	US>M	P>M	US>P	—	—
C3: Dutifulness						
15	—	—	—	US<P	—	P>M
45	—	—	—	—	—	P<M
105	—	—	—	—	US>M	P<M
225	—	—	—	US>P	—	P<M
C4: Achievement-Striving						
20	—	—	P>M	US<P	—	—
80	—	—	—	—	US<M	P<M
110	—	—	—	—	US>M	—
140	—	—	—	—	US<M	P<M
200	—	—	—	—	—	P>M
230	—	—	—	—	—	P>M
C5: Self-Discipline						
25	—	—	—	US<P	—	—
55	—	—	—	—	US<M	—
85	—	—	—	US>P	—	—
235	—	—	—	—	US>M	P>M
C6: Deliberation						
60	—	—	—	US<P	—	—
90	—	—	—	—	US<M	—
120	—	—	—	US<P	—	P>M
150	—	—	—	—	US<M	—
210	—	—	—	—	US>M	P>M
240	—	—	—	US<P	—	P>M

Note. NEO-PI-R = Revised NEO Personality Inventory; DIF = differential item functioning; Phil. = Philippines; Mex. = Mexico. Dashes indicate that there was no loading or intercept DIF for the item in a given cultural comparison.

Received June 22, 2010
Revision received April 11, 2011
Accepted April 27, 2011 ■