Journal of Research in Personality 42 (2008) 1524–1536

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Testing the measurement equivalence of personality adjective items across cultures ☆

Christopher D. Nye [a,*], Brent W. Roberts [a], Gerard Saucier [b], Xinyue Zhou [c]

[a] *Department of Psychology, University of Illinois at Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820, USA*
[b] *Department of Psychology, University of Oregon, 1227, Eugene, OR 97403, USA*
[c] *Department of Psychology, Sun Yat Sen University, Guangzhou 510275, China*

ARTICLE INFO

Article history:
Available online 26 July 2008

Keywords:
Measurement equivalence
Cross-cultural personality

ABSTRACT

Although previous research has examined cross-cultural differences in personality, many of these studies neglected to first establish that the measures being used were equivalent in meaning across cultures. Using samples of Chinese, Greek, and American respondents, the measurement equivalence of the Big Five Mini-Markers [Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality assessment, 63*, 506–516] was assessed using confirmatory factor analysis. The results indicate that all of the scales demonstrate configural invariance, but fail to show metric or scalar invariance. Several adjectives from these scales were found to exhibit bias at the item-level. The practical implications of these results are discussed and future research is suggested.

1. Introduction

The cross-cultural generalizability of personality constructs has been a long standing interest of personality researchers. In fact, Triandis (1997) stated that the study of personality is the oldest facet of cross-cultural research. As researchers attempt to determine the nature of personality traits, this form of research is particularly important for assessing the generalizability of individual differences. Some have not only questioned the cross-cultural generalizability of the structure of personality traits but the concept of traits in general (Markus, 2004). Needless to say, if the meaning and structure of personality traits does not generalize across cultures then the idea of stable individual differences would be undermined.

Cross-cultural studies shed light on whether personality traits are universal constructs that apply to all people (McCrae & Costa, 1997) or a culturally based phenomenon that varies in different regions around the world (Shweder & Sullivan, 1993). If traits are a human universal, behavior in every culture can be explained using common constructs and taxonomies. This common framework could facilitate comparisons of personality traits across cultures and provide further information for developing a comprehensive theory of the origin and development of these constructs (Church & Lonner, 1998).

Although there has been much research on the similarity of personality factor structures across cultures (e.g., Saucier, 2003) and on the appropriateness of using etic measures of personality (e.g., Katigbak, Church, & Akamine, 1996), no previous research has performed a thorough assessment of measurement equivalence. Equivalent measurement is an important first step to cross-cultural comparisons and a psychometric necessity for making claims of the generalizability of the

☆ An earlier version of this paper was presented at the Society for Industrial and Organizational Psychology Annual Conference, San Francisco, California, April, 2008.

* Corresponding author.
E-mail address: cnye2@cyrus.psych.uiuc.edu (C.D. Nye).

0092-6566/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jrp.2008.07.004

meaning and structure of personality traits (Church, 2001). If people in different cultures do not use similar words in equivalent fashion, then any inferences concerning personality traits and personality trait structures across cultures are tentative at best. Moreover, ignoring measurement equivalence may produce misleading comparisons of means, differences, and/or factor structures. Unfortunately, the bulk of previous research has primarily used exploratory methods to test equivalence (Church & Burke, 1994). In this context, confirmatory methods are more appropriate because they are less reliant on "heuristic-guided decisions" (Steel et al., 2006). However, confirmatory methods have yet to be used to test for measurement equivalence across cultures. In this article, we use Mean and Covariance Structures (MACS) techniques (Little, 1997) to examine the equivalence of a set of Big Five adjectives used across three distinctly different cultures: American, Greek, and Chinese.

1.1. Personality traits across cultures

Two primary methods have dominated research on the comparability of personality traits across cultures. The first method reflects an emic approach to research in which the culturally specific derivation of personality traits is the primary focus. Typically, these studies have examined the structure of personality trait lexicons derived from indigenous dictionaries or word lists that describe personality traits (Saucier & Goldberg, 2001). The second method has assumed an etic approach in which scores on a test developed in one culture, typically the U.S., are compared across different cultures (e.g., McCrae & Costa, 1997). As measurement equivalence techniques are only applicable to the etic study of personality, we review this literature before discussing measurement equivalence techniques in further detail.

To this point, a majority of cross-cultural personality research has focused on testing the generalizability of factor structures found in U.S. samples, typically the Five Factor Model (FFM). For example, McCrae, Costa, and Yik (1996) compared Chinese and American responses on the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992). According to their interpretation, the five-factor structure fit well in the Chinese culture. Also using the NEO-PI-R, McCrae and Costa (1997) found congruence coefficients between Chinese and American factor structures of .92 and higher. Similar results have been found in other languages using similar methods. These languages include German, Portuguese, Hebrew, Korean, Japanese (McCrae & Costa, 1997), French, Filipino (McCrae, Costa, del Pilar, Rolland, & Parker, 1998), Finnish, and Polish (Paunonen, Jackson, Trzebinski, & Forsterling, 1992).

Several large-scale studies have also supported the generalizability of personality using a number of measures. In the first such study, Barrett and Eysenck (1984) examined the factor structure of the Eysenck Personality Questionnaire (EPQ) across 25 separate cultures. Using exploratory factor analyses and a Promax rotation, they found that the same four factors (i.e. extraversion, neuroticism, psychoticism, and social desirability) could be extracted in all of the cultures. A later extension of this work (Lynn & Martin, 1995) replicated these results in a larger sample of 37 countries.

Similarly, McCrae and Terracciano (2005) used the NEO-PI-R questionnaire to assess the universality of the FFM across 50 cultures. Again using exploratory factor analyses, these authors found that, with the exception of several African countries, the FFM was replicated in each culture. Even in the dissimilar African cultures, only the openness factor was less clearly replicated.

In the largest study to date, Schmitt et al. (2007) compared the five-factor structure of the Big Five Inventory (BFI) across 56 nations. Although a six factor structure was discernable, these authors suggested that the sixth factor was superfluous and substantively unimportant. Moreover, this study showed that the five factors were virtually identical to the five U.S. factors, with a total congruence coefficient of .98.

The congruence of factors is an important first step in establishing the equivalence of measures across cultures, but it is not the last step. As we describe below, several additional levels of equivalence need to be established before comparing scores across cultures. Despite this necessity, several studies have made comparisons of mean-level differences across cultures without addressing these forms of equivalence. For example, McCrae (2001) attempted to test cross-cultural differences in mean levels of personality traits. Noting that three-quarters of the 84 sampled countries showed mean levels similar to those of American samples, McCrae claimed that there is more variation within cultures than between cultures. In contrast, McCrae et al. (1996) found that Chinese respondents were lower than Americans in extraversion and conscientiousness, but higher in neuroticism and agreeableness. Similar results were also found by Yang (1986).

Testing the validity of cultural stereotypes, Terracciano et al. (2005) used adjectival items to assess the national character of 49 separate cultures. In doing so, these authors found mean-level differences in the personality profiles of these cultures. Similarly, Schmitt et al. (2007) provided trait profiles of 56 nations and concluded that there were significant differences among nations on all five of the factors assessed with effect sizes ranging from small to moderate. Specifically, the most extraverted people tended to live in Serbia or Croatia, whereas the most agreeable people lived in Jordan and the Democratic Republic of Congo. The highest scoring nations on the conscientiousness scale were the Democratic Republic of Congo and Slovenia. Japan and Argentina were the highest on neuroticism while respondents from Chile and Belgium rated themselves highest on openness to experience. Despite the intuitive appeal and the comprehensive nature of such large-scale studies, without first establishing measurement equivalence, these differences may be inaccurately attributed to true group differences.

In summary, the argument for the comparability of personality structures and means across cultures is based on a limited evaluation, since no studies have tested the equivalence of personality traits using an appropriate methodology. Given the differences in meaning and content of various factors that emerge in emic lexical studies (Saucier & Goldberg, 2001), one might expect differences in the meaning of single trait adjectives across cultures under the assumption that the definitions and usages of specific words may be culturally specific. Therefore, tests of measurement equivalence are needed to ensure the comparability of these types of scales.

1.2. Testing the measurement equivalence of personality traits across cultures

In recognition that mean differences may be the result of a poor translation, McCrae, Yik, Trapnell, Bond, and Paulhus (1998) used a *t*-test to test for differential item functioning (DIF) between North American and Chinese students. However, this method of assessment has been shown to be inadequate for assessing measurement equivalence (Drasgow, 1987). More specifically, this method does not distinguish between true score and observed score differences. A respondent's true score on a test is their actual level of the trait that a personality measure is attempting to quantify. In contrast, their observed score is their actual score on a test. Comparing means ignores the role of error in the observed score, error which may be caused by an inappropriate measurement of the latent trait. For these reasons, other methods of assessing the equivalence of means are more appropriately used.

There are several ways in which measurement equivalence can be assessed. The primary method used by personality researchers to establish the equivalence of factor structures across cultures has been exploratory factor analyses (EFA) followed by an evaluation of the factor congruence across countries (van de Vijver & Leung, 2001). However, several recent articles have suggested using a confirmatory factor analytic (CFA) approach to assessing measurement equivalence (Cheung & Rensvold, 1999, 2000; Little, 1997; Vandenburg & Lance, 2000). Known as Mean and Covariance Structures (MACS) analysis, this method has been described as "ideally suited to establish construct comparability and, at the same time, detect possible between group differences" (Little, 1997, p. 54). These methods are appealing for several reasons. First, they allow for confirmatory tests of the relationships between variables at the construct level (Little, 1997). Second, measurement error within the factors is estimated and removed providing parameter estimates that are theoretically free of error (Ullman, 2006). Third, MACS is functionally equivalent to an item-response theory approach, while not imposing such formidable power demands to test the model (e.g., needing thousands of participants; Stark, Chernyshenko, & Drasgow, 2006). Fourth, MACS analyses are capable of more advanced assessments of equivalence, including tests of the invariance of means and differences across cultures (Little, 2000). In fact, this is one of the primary advantages of MACS over the more traditional EFA approach. Finally, once model parameters have been estimated, the quality of the specified model can be assessed using a combination of several fit statistics that describe the appropriateness of the model given the data.

As a result of these advantages, an increasing number of studies have begun to use these techniques to ensure the equivalence of factor structures across groups. For example, Oishi (2007) used MACS to examine the equivalence of a measure of subjective well-being in a Chinese sample. The results of this study showed that items assessing the contribution of personal accomplishments to life satisfaction were biased in the Chinese sample. MACS procedures have also been used to find non-equivalence in measures of individualism and collectivism (Chirkov, Ryan, Kim, & Kaplan, 2003).

Using MACS analyses for assessing measurement invariance consists of several steps, each one a stronger test of invariance than the previous. In a comprehensive review of the literature, Vandenburg and Lance (2000) found that the number and order of the recommended steps for assessing measurement invariance differed widely. Following Steenkamp and Baumgartner's (1998) recommendation to link the forms of invariance assessed to the purposes of the study, we focus on three types of measurement invariance: configural invariance, metric invariance, and scalar invariance.

Configural invariance is frequently recommended as the first step in assessing measurement equivalence (Vandenburg & Lance, 2000). This test assesses the extent to which the number of factors and the pattern of factor loadings is equivalent across groups (Horn & McArdle, 1992). Stated differently, configural invariance examines whether the factor structure of the measure holds in each of the groups. Failure to accept the null hypothesis that the pattern of factor loadings is equivalent across groups indicates that no further tests of invariance are needed (Cheung & Rensvold, 2001; Vandenburg, 2002; Vandenburg & Lance, 2000). If configural invariance is found, a test for metric invariance would follow and the configural model would be used as a baseline for comparing the fit of the constrained models.

The next step, metric invariance, constrains the factor loadings of the model for one group to be equivalent to the corresponding factor loadings in the other groups (i.e. $\boldsymbol{\lambda}^g = \boldsymbol{\lambda}^{g'}$). The interpretation of the factor loadings indicates that support for this type of invariance allows meaningful comparisons of differences across groups.

In the last step, scalar invariance tests the appropriateness of comparing means by constraining the intercepts (and the factor loadings) of the items to be equal across groups (i.e. $\boldsymbol{\tau}^g = \boldsymbol{\tau}^{g'}$ as well as $\boldsymbol{\Lambda}^g = \boldsymbol{\Lambda}^{g'}$). This test has been interpreted by some as an assessment of systematic response bias (e.g., leniency; Vandenburg & Lance, 2000). Here, individuals in one group may score differently on an item solely because of the way they conceptualize the behavior being assessed and not because of true mean differences. In other words, although personality exists in different cultures, the way it is measured may affect the meaning of traits when compared across groups. Therefore, before mean-level comparisons can be interpreted as reflecting actual cultural differences, it must be shown that individuals with the same level of the latent trait are equally likely to endorse an item, regardless of group membership. As described previously, past personality research has neglected to assess this type of invariance appropriately prior to making mean-level comparisons.

Although metric and scalar invariance are typically assessed separately, Stark et al. (2006) suggest several reasons for assessing both simultaneously. First, a separate assessment of both types of equivalence may be unnecessarily cumbersome. The results of their simulations indicate that the CFA approach is equally sensitive to bias when loadings and intercepts are assessed simultaneously. Second, separate assessments increase the probability of a Type I error. Finally, sequential tests may propagate errors in invariance detection that occur earlier in the process.

1.3. The current study

This article examines the measurement equivalence of one measure of personality in several cultures. For our first study, the measurement equivalence of the Big Five Mini-Marker scales (i.e. Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Intellect) will be assessed across two English-speaking samples. Despite the benefits of adjective items, this method of assessing personality has been criticized for lacking context and specificity (Briggs, 1992). Definitional ambiguities may lead independent groups of respondents, even within a single culture, to rate the same items differently. Therefore, before testing cross-cultural differences, it is important to show that these items function equivalently within a single culture. Therefore, in Study 1 we assess measurement equivalence across two different groups within a single culture (i.e. United States). With these analyses, we aim to establish a context within which to interpret the cross-cultural equivalence results. Items that are not measured equivalently within a single culture are also likely to not be measured equivalently across cultures. This information may shed light on the potential source of any bias that is detected cross-culturally (i.e. whether the problem resides with the item itself or cultural idiosyncrasies).

In Study 2, we test the measurement equivalence of the Mini-Markers across three cultures. Again, biased scales will be examined at the item-level. Mean-level comparisons between the cultures assessed in this study will also be examined for significant differences using both the biased and unbiased scales. These analyses are intended to function as an illustration of the potential effects of comparing biased scales across cultures.

Past research (Saucier, Georgiades, Tsaousis, & Goldberg, 2005) has suggested that the phylogenetic or geographic relationship between languages may be responsible for the similarities and differences observed in previous cross-cultural research. Stated differently, languages from geographically similar regions of the world or with a relatively similar developmental history are more likely to be equivalent to each other than to more distant languages. Therefore, the current study will assess personality in several cultures (i.e. American, Chinese, and Greek) that are different in phylogeny, geography, culture, and religion. We believe this will provide a more stringent test of the comparability of personality across cultures.

Previous research (Church & Burke, 1994; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996) has also suggested that the CFA approach is inappropriate for studies of the FFM because of a mismatch between the assumptions of the CFA and trait structure models. However, research by Roberts, Chernyshenko, Stark, and Goldberg (2005) has successfully applied the CFA model at the more unidimensional factor level. Therefore, the current study does not attempt to test the FFM in its entirety. Instead, we will assess each of the Big Five traits separately using the CFA approach.

Importantly, the present research does not devalue the role of EFA in personality research. In fact, EFA techniques are appropriate under many circumstances. For example, emic studies attempt to find the unique best structure for the target language and compare this indigenous structure to previously identified models. For these studies, the purpose is not necessarily to force the data into a previously determined factor structure. Therefore, EFA are well-suited for addressing this research question whereas CFA are not. In contrast, measurement equivalence addresses the applicability of etic measures with good scientific rationales behind them. This question is necessarily “downstream” from the emphasis of emic studies of personality and is not addressed sufficiently by EFA techniques. Therefore, although we advocate the use of CFA methodology for assessing measurement equivalence, EFA play an important role in the study of personality.

2. Study 1

2.1. Methods

2.1.1. Samples

For our within-culture comparisons, the first American sample consisted of 727 undergraduate students at two large Midwestern universities. The sample was comprised of 388 women and 339 men and the mean age was 19.27. These individuals were compared to a second sample of 712 undergraduate students from another large Midwestern university. This sample included 292 men and 399 women and the mean age was 19.16.

2.1.2. Measure

The measure used in this study was a 40-item adjective measure of the five-factor structure developed by Saucier (1994). These “Mini-Markers” are a reduced set of the 100-markers developed by Goldberg (1990). The scale includes eight items for each of the five factors and response options range from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*).

2.1.3. Analyses

In order to assess the extent of measurement equivalence, we conducted multi-group confirmatory factor analyses on group covariance matrices using LISREL VIII (Jöreskog & Sörbom, 1993). Equivalence was assessed in two steps. First, the configural invariance of each scale was examined. Next, for scales found to demonstrate configural invariance, the factor loadings and intercepts were simultaneously constrained to be equal in each sample, thereby testing metric and scalar equivalence. If bias was found in the means and differences of the groups, the loadings and intercepts were simultaneously constrained across groups for a single item at a time. This method indicated which items were responsible for bias among the groups and is similar to the approach used by Oishi (2007). Stark et al. (2006) showed that constraining a single item at a time is preferable to detecting bias by constraining all items across groups and systematically freeing single items. In this case, the detection of non-equivalence may be adversely affected if the baseline model contains a number of biased items.

In CFA, it is common to set the variance of a latent factor by constraining the loading of a single item (hereafter referred to as the referent) to unity. However, detecting invariance may be problematic if the chosen item is biased against one or more of the groups examined. To eliminate this problem, the present study applied a systematic procedure proposed by Cheung and Rensvold (1999). Here, items were considered unbiased only if they were invariant when tested using all possible items in the sub-scale as referents. For example, if a model has four indicators and λ_{21} is constrained to be equivalent across groups, three separate models would be tested with λ_{11}, λ_{31}, or λ_{41} constrained to unity. Item 2 would be considered invariant only if all three models fit the data well. If an item was shown to be unbiased in the current research, it was used as the referent item for subsequent models.

The fit of each of the MGCFA models was assessed using traditional model fit statistics including RMSEA, CFI, and NNFI. Due to the well-known sensitivity of the χ^2 statistic to sample size and model complexity, models were judged to be biased if the RMSEA of the constrained model was outside of the 90% confidence interval for the RMSEA of the baseline model. In addition, models were also evaluated by examining their absolute fit relative to the generally accepted rules of thumb (Hu & Bentler, 1999) for these indices.

2.2. *Results*

2.2.1. *Factor structures*

As the first step in our analyses, we examined the factor structures of each Big Five trait in the American sample of 727 students using confirmatory factor analyses. As shown in Table 1, the one-factor models did not fit the American data well. Therefore, we took the following steps to determine the appropriate factor structures to use for assessing configural invariance. First, to determine the patterns of factor loadings for these models, we used exploratory factor analyses (EFA) with a Promax rotation. The results of these analyses are shown in Table 2. Second, the models suggested by the EFA were tested in confirmatory models. These analyses are discussed next.

For extraversion, EFA indicated a two-factor structure defined by positive and negative terms, respectively, was appropriate. We first tested a strict two-factor CFA model with the adjectives bashful, shy, quiet, and withdrawn loading on a second factor. Although this model fit moderately well (RMSEA = .095, CFI = .96, CFI = .91), the RMSEA was higher than typically recommended (Hu & Bentler, 1999). The standardized residuals suggested that improvements in fit might be gained by estimating secondary loadings for bashful, shy, quiet, and withdrawn on the first factor. This model fit well (RMSEA = .075, CFI = .97, CFI = .98). Here, the deviation from a one-factor structure was clearly caused by method variance.

EFA suggested a similar model for the agreeableness factor with two factors largely defined by positive and negative terms. Again, a strict two-factor model with no cross-loadings fit the data poorly (RMSEA = .12, NNFI = .88, CFI = .91), so we estimated a new model with secondary loadings for harsh, unsympathetic, and sympathetic. This model fit substantially better (RMSEA = .059, NNFI = .97, CFI = .98).

For conscientiousness, the EFA solution showed a compelling two-factor structure. The adjectives organized, efficient, systematic, practical, and inefficient loaded on the first factor while the adjectives sloppy, careless, and disorganized loaded on a second factor. In the CFA, this model fit the American data well without specifying cross-loadings (RMSEA = .066, NNFI = .93, CFI = .96). Interestingly, conscientiousness was one of the only factors modeled well by two distinct, yet correlated, factors. However, like extraversion and agreeableness, these two factors were largely defined by positive and negative items.

A two-factor model was suggested again for the neuroticism scale. However, these factors appear to be more substantive, largely because the two positive items (unenvious and relaxed) loaded on separate factors. The first factor was focused on envy and jealousy whereas the second factor was more indicative of temperament. This model also fit the data in the American sample well without cross-loadings (RMSEA = .064, NNFI = .96, CFI = .97).

Table 1
Results of the one-factor confirmatory factor analyses

	RMSEA	NNFI	CFI	AGFI	SRMR
Extraversion	.120	.93	.95	.87	.06
Agreeableness	.120	.86	.90	.86	.07
Conscientiousness	.220	.25	.46	.65	.16
Neuroticism	.130	.84	.88	.85	.08
Intellect	.170	.71	.80	.75	.12

Table 2
Promax solutions for the exploratory factor analyses of the American student sample

Item	Factor loadings	
	1st	2nd
Extraversion		
Talkative	−.57	**.73**
Extroverted	−.51	**.62**
Bold	−.51	**.55**
Energetic	−.30	**.57**
Shy	**.89**	−.60
Quiet	**.80**	−.63
Bashful	**.65**	−.38
Withdrawn	**.52**	−.48
Agreeableness		
Sympathetic	**.80**	−.35
Warm	**.55**	−.46
Kind	**.54**	−.47
Unsympathetic	−**.67**	.41
Cooperative	.37	−**.41**
Cold	−.41	**.61**
Rude	−.38	**.62**
Harsh	−.30	**.69**
Conscientiousness		
Organized	.03	**.60**
Efficient	−.05	**.73**
Systematic	−.02	**.36**
Practical	−.01	**.40**
Inefficient	−.04	−**.53**
Sloppy	**.82**	.01
Disorganized	**.75**	.01
Careless	**.62**	−.02
Neuroticism		
Unenvious	−**.58**	−.30
Fretful	**.42**	.39
Envious	**.80**	.39
Jealous	**.83**	.43
Relaxed	−.25	−**.26**
Temperamental	.29	**.64**
Moody	.38	**.73**
Touchy	.32	**.48**
Intellect		
Creative	**.87**	.36
Imaginative	**.69**	.37
Uncreative	−**.81**	−.33
Philosophical	.26	**.56**
Intellectual	.20	**.52**
Complex	.19	**.57**
Deep	.36	**.64**
Unintellectual	−.31	−**.39**

Note: Bold values indicate the highest factor loading for each item.

Finally, the EFA also implied a two-factor structure for intellect. Although a two-factor CFA model was estimated with uncreative, imaginative, and creative loading on the second factor, the model did not fit well (RMSEA = .10, NNFI = .88, CFI = .93). The modification indices indicated that substantial improvements in fit could be gained with two additional constraints. First, the adjective imaginative was constrained to have a secondary loading on the first factor. Second, the unique variance in the intellectual/unintellectual pair was allowed to correlate. Given that these adjectives are antonyms of each other, this constraint is justified. The resulting model fit the data well (RMSEA = .080, NNFI = .93, CFI = .96).

2.2.2. Measurement invariance

Next, we tested the three different forms of measurement equivalence across the two U.S. samples. Table 3 summarizes the results from the multi-group comparisons. Configural equivalence was tested by comparing the factor structures developed in the initial exploratory and confirmatory analyses across the two samples. As shown in the table, all five of the factors demonstrated configural invariance. Thus, these more refined factor structures, which in most cases reflected methodological factors, replicated across the two groups. Next, metric and scalar equivalence were tested by constraining the item load-

Table 3
Fit statistics for MGCFA models across two english-speaking samples

	χ^2	df	RMSEA	90% CI for RMSEA	NNFI	CFI
Extraversion	218.43	48	.079	.068 to .089	.96	.98
Metric/scalar model	145.98	47	.064	.052 to .075	.98	.98
Agreeableness	276.33	48	.091	.081 to .100	.94	.96
Metric/scalar model	466.26	78	.090	.082 to .099	.93	.94
Conscientiousness	330.22	57	.086	.077 to .094	.94	.96
Metric/scalar model	190.96	38	.079	.069 to .091	.93	.95
Neuroticism	178.73	47	.070	.060 to .081	.95	.97
Metric/scalar model	157.78	48	.068	.057 to .079	.95	.96
Intellect	202.85	45	.078	.067 to .089	.95	.97
Metric/scalar model	189.26	46	.074	.063 to .085	.94	.95

Note: Values listed for the five trait names represent the fit statistics for the configural models.

ings and intercepts. This multi-group comparison showed no differential item functioning at the metric or scalar level. These results suggest that any differential item functioning observed in Study 2 is likely to be a result of cultural factors.

3. Study 2

3.1. Methods

3.1.1. Sample and measure

For the cross-cultural analyses, we compared the American sample of 727 undergraduate students from Study 1 to Greek and Chinese samples. The Greek sample consisted of 991 university students from several Greek universities and included 224 men and 751 women (16 did not indicate their gender). The Chinese sample was comprised of 433 university students from a large university in Shanghai. This sample contained approximately 49% women and 51% men and was a subset of the dataset analyzed by Zhou, Saucier, Gao, and Liu (in press). All participants completed the Mini-Markers Scale used in Study 1 in their native language and the analyses proceeded in a similar manner with the American, Greek, and Chinese cultures as the groups. Translations were carried out by native speakers in each of the cultures and independently confirmed by several additional translators. Again, models were evaluated relative to the 90% confidence interval for the RMSEA of their baseline model and their absolute level of fit.

3.2. Results

3.2.1. Configural invariance

First, we tested the generalizability of the factor structures found in the American sample to the Chinese and Greek cultures using multi-group CFAs. The results of these analyses for all five of the factors are presented in Table 4. Results suggest that the two-factor models fit well in all three cultures. The RMSEA statistics were all .08 or below, and with the exception of

Table 4
Fit statistics for the tests of measurement equivalence across cultures

	χ^2	df	RMSEA	90% CI for RMSEA	NNFI	CFI
Extraversion						
Configural invariance	260.32	48	.079	.069 to .088	.93	.96
Metric/scalar invariance	851.93	78	.120	.110 to .120	.83	.84
Agreeableness						
Configural invariance	275.31	48	.081	.071 to .090	.91	.95
Metric/scalar invariance	849.10	78	.120	.110 to .130	.81	.83
Conscientiousness						
Configural invariance	287.16	57	.077	.069 to .086	.91	.94
Metric/scalar invariance	2093.25	81	.180	.170 to .190	.47	.48
Neuroticism						
Configural invariance	318.04	57	.079	.070 to .088	.89	.93
Metric/scalar invariance	2193.72	81	.200	.190 to .210	.39	.41
Intellect						
Configural invariance	301.86	54	.081	.072 to .090	.91	.94
Metric/scalar invariance	1643.66	80	.170	.160 to .180	.60	.62

the neuroticism scale (NNFI = .89), all of the NNFI and CFI statistics were above .90. Therefore, all of the scales demonstrated configural invariance across groups for each of the Big Five scales.

3.2.2. Metric and scalar equivalence

After examining the equivalence of the factor structures across cultures, we simultaneously tested for metric and scalar equivalence. The fit statistics in Table 4 show that none of the Big Five factors demonstrated these forms of equivalence across the cultures examined. In each case, the resulting RMSEA was larger than the baseline model (the two-factor configural model) and was outside of the 90% confidence interval.

In order to determine which items were causing the inequality in each of the factors, we assessed measurement equivalence at the item-level. Table 5 shows that several items were found to be problematic for each of the five personality traits when the resulting RMSEA was compared to the baseline model. Moreover, the number of items that exhibited DIF varied for each of the factors. In the case of extraversion, the adjectives bold and bashful were not equivalent across cultures. For the agreeableness scale, kind, unsympathetic, and cooperative showed differential item functioning. In the case of conscientious-

Table 5
Item-level fit statistics for cultural MGCFAs

	χ^2	df	RMSEA	90% CI for RMSEA	NNFI	CFI
Extraversion	260.32	48	.079	.069 to .088	.93	.96
Talkative	328.00	52	.086	.077 to .095	.91	.94
Extroverted[a]	309.25	52	.082	.073 to .091	.92	.95
Bold	**406.18**	**52**	**.097**	**.088 to .110**	**.88**	**.93**
Energetic	308.17	52	.082	.073 to .091	.92	.95
Shy[a]	304.12	52	.083	.074 to .092	.91	.94
Quiet	311.79	54	.082	.073 to .090	.92	.95
Bashful	**408.82**	**54**	**.095**	**.087 to .099**	**.89**	**.93**
Withdrawn	343.49	54	.086	.077 to .095	.91	.94
Agreeableness	275.31	48	.081	.071 to .090	.91	.95
Sympathetic[a]	337.04	55	.085	.077 to .094	.90	.94
Warm	339.88	55	.084	.075 to .092	.90	.94
Kind	**422.72**	**55**	**.094**	**.086 to .100**	**.87**	**.92**
Unsympathetic	**470.41**	**55**	**.100**	**.093 to .110**	**.86**	**.91**
Cooperative	**417.41**	**57**	**.092**	**.084 to .100**	**.88**	**.92**
Cold[a]	337.04	55	.085	.077 to .094	.90	.94
Rude	355.66	55	.087	.079 to .096	.90	.93
Harsh	327.38	57	.081	.073 to .090	.91	.94
Conscientiousness	287.16	57	.077	.069 to .086	.91	.94
Organized	329.68	61	.080	.072 to .089	.91	.93
Efficient	369.01	61	.085	.077 to .093	.89	.92
Systematic	322.68	61	.079	.071 to .087	.91	.93
Practical[a]	369.01	61	.085	.077 to .093	.89	.92
Inefficient	**816.82**	**61**	**.130**	**.120 to .140**	**.73**	**.81**
Sloppy	**893.64**	**61**	**.130**	**.120 to .140**	**.72**	**.80**
Disorganized	**720.68**	**61**	**.120**	**.110 to .130**	**.77**	**.83**
Careless	**893.34**	**61**	**.130**	**.120 to .140**	**.72**	**.80**
Neuroticism	318.04	57	.079	.070 to .088	.89	.93
Unenvious	**655.92**	**61**	**.120**	**.110 to .130**	**.77**	**.83**
Fretful	**783.32**	**61**	**.120**	**.110 to .130**	**.72**	**.80**
Envious	**1122.17**	**61**	**.140**	**.130 to .150**	**.59**	**.71**
Jealous	**1122.17**	**61**	**.140**	**.130 to .150**	**.59**	**.71**
Relaxed	**612.64**	**61**	**.110**	**.100 to .120**	**.79**	**.85**
Temperamental	**499.85**	**61**	**.097**	**.089 to .110**	**.83**	**.88**
Moody	**623.21**	**61**	**.110**	**.100 to .120**	**.78**	**.84**
Touchy	**623.21**	**61**	**.110**	**.100 to .120**	**.78**	**.84**
Intellect	301.86	54	.081	.072 to .090	.91	.94
Creative	**827.42**	**58**	**.130**	**.120 to .140**	**.73**	**.81**
Imaginative	**544.96**	**60**	**.110**	**.099 to .120**	**.83**	**.88**
Uncreative	**852.26**	**58**	**.130**	**.120 to .140**	**.72**	**.81**
Philosophical	**501.86**	**58**	**.100**	**.094 to .110**	**.84**	**.89**
Intellectual	**537.64**	**58**	**.110**	**.099 to .120**	**.83**	**.88**
Complex	**421.96**	**58**	**.094**	**.086 to .100**	**.87**	**.91**
Deep	**537.64**	**58**	**.110**	**.099 to .120**	**.83**	**.88**
Unintellectual	**495.89**	**58**	**.100**	**.092 to .110**	**.84**	**.89**

Note: For factors without an unbiased referent, the fit statistics for the poorest fitting model are shown. Therefore, many of the values are similar. However, they are not directly comparable because misfit is determined relative to the bias in the referent item. In this case, reporting the worst fitting model only provides a rough illustration of the amount of DIF present in the item.

[a] Indicates unbiased referent items.

ness, inefficient, sloppy, disorganized, and careless were biased. In contrast, all of the items in the neuroticism and intellect scales were biased at the metric and scalar level. Importantly, the non-equivalence appeared to be pervasive. It was not related to the underlying configural structure of any of the five scales and, with the exception of conscientiousness, was not related to positively or negatively termed items. Moreover, different numbers of items were biased in each of the Big Five scales.

To more fully examine the non-equivalence, we also tested MGCFA models with only the factor loadings constrained to be equivalent. This helped to determine whether the bias was in the loadings or the intercepts of each item. Ten items showed bias at both levels (shown in bold) while 15 items showed only scalar bias (shown italicized bold). In addition, the group-level fit statistics (i.e. SRMR and GFI) confirmed that the non-equivalence was pervasive in all sample comparisons (i.e. American vs. Greek and American vs. Chinese) and not restricted to a single culture.

The results presented here show equivalent configural structures across cultures, but problematic bias at the level of metric and scalar equivalence. This bias could strongly affect comparisons of means and standard deviations across cultures. In order to illustrate how this bias might affect cross-cultural comparisons, we tested mean-level differences in the Big Five scales across cultures. We first show the differences that would result without considering measurement equivalence and then after discarding the items that showed differential item functioning (DIF). Neuroticism and intellect were necessarily excluded from these analyses because all of the items in these scales were biased.

Table 6 shows the means and standard deviations for (1) the overall factor scales with all items included in the sum, (2) a scale score for each Big Five dimension using only items that did not show bias, and (3) the items themselves (for illustrative purposes, the means and standard deviations of the overall scale scores and the items are reported for neuroticism and intellect). After a Bonferroni correction for multiple pairwise comparisons, several interesting and statistically significant results were found when mean differences were examined for the full factor scales.

First, when using the full set of adjectives, the American respondents showed significantly higher levels of extraversion than the Greek ($d = .25$) or Chinese samples ($d = .41$). In turn, the Greek sample rated themselves higher than the Chinese on total extraversion ($d = .20$). In contrast, when the unbiased items were used to measure extraversion there were no statistically significant differences between American and Greek samples ($p = .11$) whereas the size of the Chinese-American difference increased ($d = .66$). Examining the item-level differences, it appears that the biased items "bold" and "bashful" both contributed contributed to the biased estimate of cultural differences in the total extraversion scores.

For agreeableness, the Greek sample scored higher than both the American ($d = .34$) and Chinese samples ($d = 1.04$). Also the American sample scored significantly higher than the Chinese sample on the biased estimate of agreeableness ($d = .65$). The unbiased estimate of agreeableness showed a similar pattern with changes in the magnitude of the differences. Specifically, the Greek sample still scored significantly higher than the American sample ($p < .05$), although the size of the effect decreased substantially ($d = .23$). In contrast, the effect size of the difference between the American and Chinese samples remained approximately equal to the biased comparison ($d = .69$). At the item-level, the only biased term that could have contributed to the Greek sample scoring higher was "unsympathetic." Thus, the Greek and American samples interpreted the meaning of this item differently.

The Greek sample also scored significantly higher than both American ($p < .05$, $d = .66$) and Chinese respondents ($p < .05$, $d = .48$) on the biased version of the conscientiousness scale and the Chinese sample scored significantly higher than the American sample, although the effect was quite small ($p < .05$, $d = .16$). However, when the conscientiousness scores were recomputed with the four unbiased items, the size of the differences between the Greek and American ($d = .24$) and Greek and Chinese samples ($d = .28$) decreased substantially. In contrast, when the biased items were removed, the difference between Chinese and American samples was reversed with Americans reporting significantly higher levels of conscientiousness than the Chinese ($p < .05$, $d = .57$). At the item-level it appears that the terms "sloppy" and "careless" contributed most to the biased estimate as the differences between the Greek and American samples were quite large for these items.

4. Overall discussion

The purpose of this research was to assess the measurement equivalence of the Big Five Mini-Markers across several cultures using statistical models that have not been fully utilized in previous cross-cultural personality research. This assessment is an absolute necessity before appropriate conclusions can be made regarding cultural differences in personality. In the absence of measurement equivalence, comparisons of measures across cultures, whether focusing on means or covariances, are suspect and possibly misleading.

In an effort to establish that any differences between cultures were not the result of the higher level of scrutiny found in confirmatory models, we first tested the comparability of the trait scales across two English-speaking samples. Consistent with the research that finds a discrepancy between exploratory and confirmatory methods when applied to personality adjectives, simple one-factor models of each trait did not fit in the American samples. However, once reasonable two-factor structures were identified, further tests showed that all five personality traits were measured equivalently when two University student samples were compared within a single culture. These results provide a baseline comparison for the second set of analyses testing the comparability of personality ratings across cultures. In light of this equivalence, differences found across cultures are less likely the result of a more stringent measurement model than actual cultural differences in the understanding and use of the trait adjectives themselves.

Table 6
Factor level means, standard deviations, and factor loadings by culture

	American (*N* = 724)	Greek (*N* = 991)	American-Greek	Chinese (*N* = 433)	American-Chinese	Chinese-Greek
			d		*d*	*d*
Extraversion	27.19 (6.12)	25.96 (4.76)	.25	24.98 (4.92)	.41	.20
Unbiased	20.56 (4.81)	20.31 (4.01)	.08	17.86 (2.76)	.66	.67
Talkative	3.68 (1.07)	3.59 (1.14)	.08	3.30 (1.08)	.35	.25
Extroverted	3.21 (1.09)	3.44 (1.25)	.19	3.12 (1.19)	.08	.25
Bold	3.42 (.95)	3.34 (1.18)	.08	2.44 (1.04)	.99	.79
Energetic	3.91 (.90)	3.76 (1.00)	.16	3.29 (1.05)	.65	.47
Shy	3.11 (1.25)	2.92 (1.30)	.14	3.13 (1.19)	.02	.17
Quiet	3.02 (1.22)	2.55 (1.28)	.38	2.82 (1.16)	.17	.21
Bashful	3.27 (1.16)	2.31 (1.06)	.88	3.01 (1.17)	.22	.65
Withdrawn	3.68 (1.09)	4.05 (1.08)	.34	3.89 (1.10)	.19	.15
Agreeableness	32.95 (4.45)	34.40 (3.71)	.34	30.03 (5.09)	.65	1.04
Unbiased	20.18 (3.16)	21.09 (2.70)	.23	19.04 (3.32)	.69	.92
Sympathetic	4.08 (.85)	4.16 (.86)	.09	3.87 (1.02)	.23	.32
Warm	4.04 (.77)	4.06 (.88)	.02	3.63 (1.06)	.47	.46
Kind	4.31 (.64)	4.30 (.73)	.01	3.69 (1.03)	.62	.74
Unsympathetic	4.36 (.86)	4.73 (.71)	.48	3.75 (1.08)	.65	1.17
Cooperative	4.15 (.74)	4.28 (.74)	.18	3.57 (1.06)	.67	.84
Cold	4.08 (1.03)	4.33 (.95)	.25	3.81 (1.08)	.26	.52
Rude	4.19 (.90)	4.61 (.70)	.52	4.16 (1.02)	.03	.55
Harsh	3.83 (1.06)	3.94 (1.19)	.09	3.57 (1.10)	.24	.31
Conscientiousness	27.20 (4.50)	30.47 (5.09)	.66	28.01 (5.06)	.16	.48
Unbiased	14.66 (2.59)	13.97 (3.15)	.24	13.12 (2.92)	.57	.28
Organized	3.54 (1.12)	3.46 (1.22)	.07	3.39 (1.08)	.13	.05
Efficient	3.90 (.85)	3.73 (.86)	.19	3.04 (1.11)	.90	.74
Systematic	3.38 (.97)	3.22 (1.21)	.17	3.21 (1.06)	.14	.01
Practical	3.87 (.83)	3.58 (1.12)	.29	3.50 (1.05)	.41	.07
Inefficient	4.00 (.89)	4.24 (1.03)	.37	3.64 (1.07)	.25	.57
Sloppy	2.87 (1.31)	4.39 (1.01)	1.34	4.52 (.91)	1.41	.13
Disorganized	2.82 (1.34)	4.15 (1.15)	1.07	3.76 (1.12)	.74	.34
Careless	2.89 (1.28)	3.70 (1.21)	.65	2.97 (1.12)	.06	.62
Neuroticism	22.36 (5.47)	20.14 (4.92)	.44	20.22 (5.01)	.41	.02
Unenvious	3.37 (1.27)	2.45 (1.42)	.67	3.25 (1.07)	.10	.60
Fretful	2.57 (1.04)	2.91 (1.30)	.35	2.26 (1.10)	.30	.52
Envious	2.75 (1.14)	1.30 (.71)	1.59	2.45 (1.11)	.27	1.35
Jealous	2.64 (1.16)	2.97 (1.37)	.26	1.71 (.95)	.86	1.01
Relaxed	2.26 (.96)	3.19 (1.19)	.84	2.80 (1.03)	.54	.35
Temperamental	2.91 (1.08)	2.60 (1.34)	.25	3.20 (1.15)	.27	.46
Moody	2.92 (1.17)	1.86 (.99)	.99	2.41 (1.10)	.45	.53
Touchy	2.99 (1.12)	2.85 (1.35)	.11	2.15 (1.11)	.75	.54
Intellect	31.09 (4.80)	28.08 (4.97)	.64	25.76 (4.49)	1.11	.42
Creative	3.77 (1.04)	3.80 (.98)	.04	3.29 (1.07)	.45	.50
Imaginative	3.86 (.98)	3.26 (1.14)	.56	3.41 (1.11)	.43	.13
Uncreative	4.00 (1.07)	4.57 (.80)	.62	3.53 (1.13)	.43	1.15
Philosophical	3.26 (1.16)	2.94 (1.34)	.25	3.19 (1.00)	.06	.20
Intellectual	4.18 (.69)	3.21 (1.16)	.98	3.52 (1.14)	.74	.27
Complex	3.77 (1.07)	2.81 (1.34)	.78	2.97 (1.27)	.70	.12
Deep	3.87 (.96)	3.29 (1.15)	.54	2.75 (1.09)	1.11	.47
Unintellectual	4.47 (.67)	4.19 (1.11)	.30	3.36 (1.19)	1.23	.73

Note: Values in parentheses are the standard deviations. Bold items are indicative of bias at both the metric and scalar levels while items that are bold and italicized were only bias at the scalar level. Compare to Table 5.

The primary goal of this study was to assess the general equivalence of trait adjectives across three different cultures: American, Greek, and Chinese. Much of the previous research on the comparability of personality structures has used exploratory factor analysis and other impressionistic techniques for determining comparability across cultures. Interestingly, the information that can be gleaned from these approaches is roughly equivalent to the first level of measurement equivalence, configural equivalence, tested in the more rigorous MACS model employed in the current study. Consistent with previous research on phrase items (McCrae & Terracciano, 2005), we found that the general structure of an adjectival scale used across these three cultures was consistent. This supports the notion that the specific traits, such as extraversion, replicate in content across cultures.

The feature unique to our study was the examination of metric and scalar equivalence, which had not been systematically applied to cross-cultural ratings of identical personality items. Contrary to the perspective that the general structure is equivalent across cultures and therefore the meaning of the scales is the same (McCrae & Costa, 1997), the present study found evidence for bias at the metric and scalar levels of analysis across cultures. Over half of the items across each of the Big Five factors were found to function differently across Chinese, Greek, and U.S. samples. Metric and scalar differences are important because bias at these levels undermines the comparability of the scale score that is often used in cross-cultural research to infer cultural differences in personality or any other psychological dimension. Together, these results suggest that although the same traits may exist in different cultures, the adjectives that describe them are used differently. Therefore, mean-level cross-cultural comparisons should be done carefully and only after ensuring measurement equivalence. Only then can these comparisons be interpreted correctly.

When we examined mean-level differences across cultures in both biased and unbiased assessments of each trait, we found dramatic effects in several circumstances. Generally speaking, the biased scales showed much larger cross-cultural differences, especially for extraversion, agreeableness, and conscientiousness. It should be noted, of course, that the meaning of the scale will inevitably change by limiting the items used to measure the construct. Nonetheless, the difference between biased and unbiased mean scores demonstrates the potentially misleading conclusions that derive from cross-cultural comparisons in which the equivalence of the measures has not been first established.

These results also illustrate the benefit of using confirmatory factor analytic methods for assessing measurement equivalence. Previous cross-cultural personality research has primarily used exploratory factor analytic methods (see Saucier & Goldberg, 2001 for a review). In general, these exploratory methods have indicated that the factor structure of personality is at least moderately similar across multiple cultures around the world at the five-factor level, though not if one applies stringent standards for factor congruence (De Raad, Perugini, & Szirmak, 1997; Saucier & Goldberg, 2001). Although the Five Factor Model was not tested in its entirety, the results of the current study provide support for the configural invariance of the five factors evaluated separately. However, exploratory methods are limited in their ability to detect other forms of bias. In addition to tests of configural invariance, CFA techniques provide the methodology for assessing bias in group-level means and differences. It was at this level that bias was found in the present study. Therefore, future assessments of cross-cultural comparisons should use the CFA approach to measurement equivalence in order to provide sufficient statistical justification for cross-cultural comparisons of means and differences.

Interestingly, neuroticism and intellect exhibited the most DIF (as determined by the number of items that were not equivalent) of any of the factors. In fact, all of the items in these two sub-scales functioned differently across cultures. Given the previous research in the domain of openness/intellect, one would expect this domain to exhibit DIF. The openness or intellect factor is the least well-defined of the Big Five. Researchers have interpreted this fifth factor at various times as culture (John & Srivastava, 1999), openness to experience (McCrae & Costa, 1985), and intellect (Goldberg, 1990). Indeed, most of the deviations from the hypothesized FFM have involved the openness factor (John & Srivastava, 1999). In contrast, the pervasiveness of bias in the neuroticism factor was surprising. Future research should examine these traits further to determine why bias was so prevalent in these items.

One possible explanation for the cross-cultural results presented here could be that single words do not sufficiently convey the meaning of these traits. Therefore, when single words are used as stimuli in a measure, individuals may have different interpretations of their meanings (Briggs, 1992). A similar explanation for the poor replication of the fifth factor in English-speaking samples has been suggested (McCrae, 1990). For this reason, some authors have suggested that phrase items may be more appropriate for cross-cultural comparisons (McCrae & Costa, 1997). While this may be the case, many phrase items rely heavily on trait adjectives for communicating the import of the item (e.g., "I act extraverted at parties"). To the extent that trait adjectives are a key source of information in personality inventory items, the lack of equivalence at the metric and scalar level should be attended to. Also, the lack of equivalence at the adjective level may provide key information for inventory developers in that they can be made aware of domains (e.g., boldness) that might lead to misinterpretations and other domains (e.g., shyness) that might not.

Despite the fact that this is one of the first studies to explicitly test the equivalence of the Big Five factors using confirmatory techniques, it still has several limitations. First, this study only focused on three particular cultures. Given the many possible countries where personality research can be conducted, future research should examine the measurement equivalence of the Mini-Markers, or other equivalent measures, in other cultures as well.

A second limitation of this study is that only adjectives were examined for equivalence. Although several studies have examined phrase items across cultures, many of these studies have also neglected to first assess measurement equivalence using the CFA approach. In contrast to the recommendations of several authors (Church & Burke, 1994; McCrae et al., 1996), the present article has shown that personality can be assessed by examining CFA results at the factor level and that these analyses may detect bias where other methods have not. An important distinction between the current study and previous attempts at CFA research is that we did not try to confirm the Big Five structure across cultures. Therefore, although this approach was sufficient for detecting bias at the construct level, it may not be appropriate for comparisons of the FFM model across cultures. Nevertheless, future research should use this framework to test phrase items for bias in other cultures.

References

Barrett, P., & Eysenck, S. (1984). The assessment of personality factors across 25 countries. *Personality and Individual Differences, 5*, 615–632.
Briggs, S. R. (1992). Assessing the five-factor model of personality description. *Journal of Personality, 60*, 253–293.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1–27.

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

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4*, 236–264.

Chirkov, V., Ryan, R. M., Kim, Y., & Kaplan, U. (2003). Differentiating autonomy from individualism and independence: A self-determination theory perspective on internalization of cultural orientations and well-being. *Journal of Personality and Social Psychology, 84*, 97–110.

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69*, 979–1006.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology, 66*, 93–114.

Church, A. T., & Lonner, W. J. (1998). The cross-cultural perspective in the study of personality: Rationale and current research. *Journal of Cross-Cultural Psychology, 29*, 32–62.

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.

De Raad, B., Perugini, M., & Szirmak, Z. (1997). In pursuit of a cross-lingual reference structure of personality traits: Comparisons among five languages. *European Journal of Personality, 11*, 167–185.

Drasgow, F. (1987). Study of measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19–29.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*, 1216–1229.

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 3*, 424–453.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York, NY, US: Guilford Press.

Jöreskog, K., & Sörbom, D. (1993). *New features in LISREL 8*. Chicago: Scientific Software International.

Katigbak, M. S., Church, A. T., & Akamine, T. X. (1996). Cross-cultural generalizability of personality dimensions: Relating indigenous and imported dimensions in two cultures. *Journal of Personality and Social Psychology, 70*, 99–114.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53–76.

Little, T. D. (2000). On the comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology, 31*, 213–219.

Lynn, R., & Martin, T. (1995). National differences for thirty-seven nations in extraversion, neuroticism, psychoticism, and economic, demographic, and other correlates. *Personality and Individual Differences, 19*, 403–406.

Markus, H. R. (2004). Culture and personality: Brief for an arranged marriage. *Journal of Research in Personality, 38*, 75–83.

McCrae, R. R. (1990). Traits and trait names: How well is openness represented in natural languages? *European Journal of Personality, 4*, 119–129.

McCrae, R. R. (2001). Trait psychology and culture: Exploring intercultural comparisons. *Journal of Personality, 69*, 819–846.

McCrae, R. R., & Costa, P. T. (1985). Updating Norman's adequate taxonomy: Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology, 49*, 710–721.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509–516.

McCrae, R. R., Costa, P. T., Jr., del Pilar, G. H., 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.

McCrae, R. R., Costa, P. T., Jr., & Yik, M. S. M. (1996). Universal aspects of Chinese personality structure. In M. H. Bond (Ed.), *The handbook of Chinese psychology* (pp. 189–207). Hong Kong: Oxford University Press.

McCrae, R. R., & Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology, 88*, 547–561.

McCrae, R. R., Yik, M. S. M., Trapnell, P. D., Bond, M. H., & Paulhus, D. L. (1998). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology, 74*, 1041–1055.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the revised NEO personality inventory: Confirmatory factor analysis versus procrustes rotation. *Journal of Personality and Social Psychology, 70*, 552–566.

Oishi, S. (2007). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality, 40*, 411–423.

Paunonen, S. V., Jackson, D. N., Trzebinski, J., & Forsterling, F. (1992). Personality structure across cultures: A multimethod evaluation. *Journal of Personality and Social Psychology, 62*, 447–456.

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

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

Saucier, G. (2003). An alternative multi-language structure for personality attributes. *European Journal of Personality, 17*, 179–205.

Saucier, G., Georgiades, S., Tsaousis, I., & Goldberg, L. R. (2005). The factor structure of Greek personality adjectives. *Journal of Personality and Social Psychology, 88*, 856–875.

Saucier, G., & Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality, 69*, 847–879.

Schmitt, D. P., Allik, J., McCrae, R. R., Benet-Martinez, V., Alcalay, L., Ault, L., et al (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.

Shweder, R. A., & Sullivan, M. A. (1993). Cultural psychology: Who needs it? *Annual Review of Psychology, 44*, 497–523.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292–1306.

Steel, R. G., Little, T. D., Ilardi, S. S., Farehand, R., Brody, G. H., & Hunter, H. G. (2006). A confirmatory comparison of the factor structure of the children's depression inventory between European American and African American youth. *Journal of Child and Family Studies, 15*, 779–794.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78–90.

Terracciano, A., Abdel-Khalek, A. M., Adam, N., Adamovová, L., Ahn, C.-k., Ahn, H.-n., et al (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310*, 96–100.

Triandis, H. C. (1997). Cross-cultural perspectives on personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 439–464). San Diego: Academic Press.

Ullman, J. B. (2006). Structural equations modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment, 87*, 35–50.

van de Vijver, F. J. R., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality, 69*, 1007–1031.

Vandenburg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139–158.

Vandenburg, 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.

Yang, K. S. (1986). Chinese personality and its change. In M. H. Bond (Ed.), *The psychology of the Chinese people* (pp. 106–170). Hong Kong: Oxford University Press.

Zhou, X., Saucier, G., Gao, D., & Liu, J. (in press). The factor structure of Chinese personality terms. *Journal of Personality*.