MEASUREMENT EQUIVALENCE OF EXTRAVERSION AND CONSCIENTIOUSNESS ITEMS AND SCALE ACROSS FOUR COUNTRIES

BY

LIWEN LIU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Advisor:

Professor Fritz Drasgow

ABSTRACT

Using IRT and MACS approaches, we examined measurement equivalence of the Extraversion and Conscientious Scales of the Big Five Inventory (BFI) between the US and China, Australia and Spain. The results showed 1-4 items of each scale to have measurement nonequivalence in each pair of comparison, although the effect sizes were not large. The mean differences in Extraversion and Conscientious between the US and the three other countries were consistent with previous findings based on cultural similarity. The study showed the feasibility of administering a personality inventory in English to non-native speakers with English proficiency as an alternative to translation.

TABLE OF CONTENTS

INTRODUCTION ……..... 1

METHOD ……...…… 12

RESULTS ……..... 17

DISCUSSION ……..... 23

CONCLUSION ……... 28

TABLES ……...……..…….. 29

REFERENCES…… 39

INTRODUCTION

Measurement equivalence (ME) has received more and more attention from researchers during the past three decades. ME holds when the relationship between observed scores and latent traits is identical across different groups (Drasgow & Kanfer, 1985). Establishing ME is a necessary first step to meaningful group comparisons, because without ME we cannot determine whether observed differences are due to true mean differences or measurement artifacts. ME has been examined in various fields of psychological and educational assessment. For example, some researchers have assessed the equivalence of computerized/internet-based and paper-and-pencil personality inventories (King & Miles, 1995; Salgado & Moscoso, 2003) and attitudinal surveys (Cole, Bedeian, & Feild, 2006; Donovan, Drasgow, & Probst, 2000; Mueller, Liebig, & Hattrup, 2007); some have examined the equivalence of performance ratings from multiple sources (Barr & Raju, 2003; Hannum, 2007; Maurer, Raju & Collins, 1998; Woehr, Sheehan, & Bennett, 2005); some have looked at the effects of time on job satisfaction (Meade, Lautenschlager, & Hecht, 2005), work ethic across cohorts (Meriac, Woehr, & Banister, 2010), or employee attitudes before and after some special event (Ryan, West, & Carr, 2003).

In addition to the contexts mentioned above, ME is also critical for cross-cultural research (e.g., Riordan & Vandenberg, 1994). Before we compare group means of respondents from different cultures, we need to confirm that all respondents use the scale format in the same way and that the scores have the same interpretation in all cultures (Mullen, 1995). There are two factors to be considered in cross-cultural studies: language and culture. People from some countries may have different cultural backgrounds although they speak the same language, such as European Spanish and Spanish-speaking Latin Americans. Also, people may share relatively

similar cultures although they speak different languages (e.g., English-speaking Canadians and French-speaking Canadians). Many cross-cultures studies have examined cultural differences for people who differ in both culture and language (e.g., English-speaking Americans vs. Chinese-speaking Chinese). However, it is important to tease apart these two factors so we can attribute observed differences to either culture or language.

In cross-cultural studies of people who speak different languages, translation is a critical process. Not only should the translation maintain the original meaning of the source language, it should also allow respondents to report in the target languages their psychological states that are equivalent to those that would be reported in the source language (Hulin, 1987). Although most cross-cultural studies have used back-translation procedures, there are still problems that are inevitable in the translation process. For example, the meaning of the items might be changed inadvertently in the process, or the items may not be equally relevant in different languages (Budgell, Raju, & Quartetti, 1995). Moreover, a successful back-translation only ensures linguistic equivalence, which is not sufficient for ME (Hulin, Drasgow, & Parsons, 1983). We also need to examine psychometric equivalence, which means two individual with the same standings on the latent trait should have the same probability of a positive response although they respond to the item in different languages (Candell & Hulin, 1986).

Approaches to Examining ME

Currently there are two approaches to examining ME: one based on item response theory (IRT) and the other based on mean and covariance structure (MACS) (e.g., Raju, Laffitte, & Byrne, 2002; Stark, Chernyshenko, & Drasgow, 2006). IRT models describe the relationship between a person's latent trait and his or her probability of a particular response to an item

(Hulin et al., 1983; Lord, 1980). In our study we used Samejima's Graded Response Model (SGRM; Samejima, 1969) for ordered categorical responses to rating scale items. The SGRM uses one discrimination parameter (the *a* parameter) and *n*-1 threshold parameters (the *b* parameters) for an item with *n* options. It is an extension of the two-parameter logistic model because each category represents a cumulative dichotomy, which is modeled by a boundary response function (BRF) that is taken as a two-parameter logistic item response function. Then a category response function (CRF) can be calculated as the difference between the two adjacent BRFs:

$$P_{ik}(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_{ik})}} - \frac{1}{1+e^{-Da_i(\theta-b_{i(k+1)})}},$$

where *i* is the item index; *k* is *k*th response option of item *i*; a_i is the discrimination parameter; and b_{ik} is threshold parameter for option *k*. $P_{ik}(\theta)$ is then the probability of endorsing option *k* of item *i* for respondents with trait level θ.

From an IRT perspective, measurement nonequivalence is referred to as differential item functioning (DIF; Raju, van der Linden, & Fleer, 1995). DIF means the probability of endorsing an item differs for respondents of a reference group and a focal group who have the same standing on the latent trait. There are generally two types of DIF: uniform DIF, which means the item favors one group of respondents across all trait levels; and nonuniform DIF, which means the item favors different groups of respondents at different trait levels. There are several methods for detecting DIF, such as Lord's χ^2 (Lord, 1980) and the Likelihood Ratio Test (Thissen, Steinberg & Wainer, 1993). In the current study we used Raju et al.'s differential functioning of items and tests (DFIT) framework, and specifically, the NCDIF (noncompensatory DIF) index to

identify DIF items. We estimated each respondent's true score on item *i* as if the respondent were from the reference group and then estimated the true score on the same item again as if the respondent were from the focal group. We found this difference of these two true scores and calculated the expectation of the squared differences. The true score differences should equal zero to guarantee ME within the DFIT framework, and the NCDIF index is compared to a cutoff value to indicate whether the item has DIF or not. When DIF occurs at the scale or test level, it is referred to as differential test functioning (DTF). Similarly, the difference in true scores at the scale/test level is calculated by summing the differences in true scores of each item. Note that it is possible to find DIF but no DTF because the DIF items may favor different groups in the same magnitude but in opposite directions and cancel out.

Our second approach, mean and covariance structure analysis (MACS), is based on the linear confirmatory factor analysis one-factor model (CFA; Jöreskog & Sörborn, 1996):

$$x = \Lambda_x \xi + \delta,$$

where x is a vector of $m \times 1$ observed variables; ξ represents the latent factor; Λ_x is an $m \times 1$ factor loading matrix that regresses the m observed variables on the factor; and δ is a $m \times 1$ vector of residuals or measurement errors.

In the CFA framework researchers have examined different degrees of ME, from relatively weaker forms such as invariant factor loading matrices to stronger forms such as invariant factor means (Vandenberg & Lance, 2000). In our study we examined three levels of ME using MACS. First, configural invariance is obtained when the number of factors and the pattern of zero and nonzero loadings are identical across groups. When configural invariance is

established, the next two forms of ME can be examined successively or simultaneously (Stark, Chernyshenko, & Drasgow, 2006). These two forms include metric invariance, which indicates equality of factor loadings, and scalar invariance which indicates equality of intercepts in addition to equality of factor loadings. For scalar invariance, we add the τ parameter to the equation:

$$x = \tau + \Lambda_x\xi + \delta,$$

where τ is the intercept for the regressions of items on the latent variables. In the basic one-factor model τ is assumed to be zero and is not estimated. The inclusion of τ allows us to examine the means of the observed scores and further the latent mean (i.e., the κ parameter) of the factor. In practice the latent mean of the reference group is set to zero while the latent mean of the focal group is expressed as the latent mean difference between the reference and focal groups (Ployhart & Oswald, 2004).

Effect Size of ME for IRT and MACS Approaches

Like most statistical tests, measurement nonequivalence in the IRT and the MACS methodologies is determined by comparing an index with the cutoff value of the corresponding significance test (e.g., the change of χ^2; see the Method section for more details). However, significance tests have been criticized for dichotomizing the decision of rejecting the null hypothesis based solely on the critical value (i.e., $p < .05$) (Kirk, 2006). For example, a *p*-value of .06 does not differ much from a *p*-value of .08 because they are both insignificant. However, a *p*-value of .04 is treated differently from a *p*-value of .06 because the former is significant but the latter is not. Moreover, statistical significance does not imply practical significance. For instance,

with a large sample size we are likely to find significant differences due to more statistical power, although the magnitude of the difference might be negligible. Therefore, it has been suggested that an effect size should be reported in addition to the significance test and various effect size indices have been developed for different types of statistical tests (e.g., Cohen's *d* for *t* test; Cohen, 1988).

For ME research, Stark, Chernyshenko, and Drasgow (2004) proposed an effect size index, d_{DTF}, for estimating the magnitude of scale-level measurement nonequivalence. In our study, we examined effect size of measurement nonequivalence at the *item* level using the indices developed by Nye (2011) and Nye and Drasgow (2011). For the IRT approach, we calculated the root mean squared difference of each item characteristic curve (ICC) between the reference group (i.e., $\mathrm{ICC}_{\mathrm{iR}}$) and the focal group (i.e., $\mathrm{ICC}_{\mathrm{iF}}$) averaged across ability level (θ). We divided this value by the pooled standard deviation of item *i* in the reference and focal groups to make the index comparable to other effect size measures (e.g., Cohen's *d*). Thus, the effect size d_{DIF} is defined as:

$$d_{DIF} = \frac{1}{SD_{iP}} \sqrt{\int [ICC_{iR}(\theta) - ICC_{iF}(\theta)]^2 f_F(\theta) d\theta}$$

where $f_F(\theta)$ is the ability density of the focal group with the mean and variance estimated from the transformed $\hat{\theta}$ distribution and SD_{iP} is the pooled standard deviation of item *i* in the reference and focal groups, which is defined as

$$SD_{iP} = \frac{(N_R - 1)SD_R + (N_F - 1)SD_F}{(N_R - 1) + (N_F - 1)}.$$

Similarly, the effect size of ME for the MACS approach used the predicted score given the latent factor ξ to replace the ICCs in the IRT approach. Thus, the effect size for MACS approach is written as

$$d_{MACS} = \frac{1}{SD_{iP}} \sqrt{\int (\hat{X}_{iR} - \hat{X}_{iF} \mid \xi)^2 f_F(\xi) d\xi}$$

where SD_{iP} is the pooled standard deviation of item *i* in the reference and focal groups and $f_F(\xi)$ is the ability density of the latent factor in the focal group.

Cross-cultural and ME Studies of Personality Inventories

Research on personality originated in the Western culture, so many studies have been etic. That is, the structure and measure of the personality construct developed and validated in the Western culture were applied in non-Western cultures as if they were universal across cultures. In spite of the potential problems of etic approach, studies have successfully replicated the five-factor structure of personality that has been found in the American culture in most other cultures (McCrae & Allik, 2002). In their book, McCrae and Allik (2002) included research from 40 cultures using the NEO-PI-R that was translated and adapted to different languages. Using the same measure in all cultures allows direct comparisons of "national personality profiles" across different countries. In general, European and American cultures were found to be higher in Extraversion and Openness to Experiences and lower in Agreeableness than Asian and African cultures (Allik & McCrae, 2004). Schmitt et al. (2007) examined the nation-level Big Five Inventory (BFI; John, Naumann, & Soto, 2008) trait profiles and found Africans more conscientious and East Asians less conscientious than people from other world regions.

Proximate cultures were more similar in personality than distant cultures (Allik & McCrae, 2004).

Although studies have shown that the five-factor model is replicable across cultures using an etic approach, it raises questions about the universality of the structure of personality: Were the concepts and measures developed in Western cultures imposed on other cultures? If each culture developed its own measure using its native language descriptors, would the five-factor model be replicated? To address this question, another line of research used the emic approach which seeks to investigate psychological construct by developing measures using the linguistic materials within each specific culture (Berry, 1969). For example, Saucier and colleagues used a lexical approach and carried out analyses with a representative set of personality descriptors within each language. Results of lexical studies showed all five personality factors in English, German and other similar languages in Northern Europe. With increasingly different languages and cultures, the five-factor structure became less constant (Saucier & Goldberg, 2001). Therefore, similarity in personality construct is more likely to be found in proximate cultures than in distant cultures, when either the etic or the emic approach is used.

Previous research has studied ME of personality inventories across languages, cultures or both using the IRT and the MACS approach. Ellis, Becker, and Kimmel (1993) conducted two DIF analyses on the English version of the Trier Personality Inventory (TPI) among American college students and the German version among German college students. Using Lord's (1980) χ^2 index they found 11 DIF items out of 120 in the first study and 7 DIF items out of 120 in the second study in which some translation errors were fixed. To further remove the effect of translation, Ellis (1995) administered the TPI to East and West Germans shortly after the

reunification and found two DIF items. Therefore, the DIF items of TPI between Americans and Germans may have resulted from both language and cultural differences.

Huang, Church and Katigbak (1997) also controlled the effect of language and examined ME of the English version of NEO Personality Inventory among Filipino and American college students. Both the IRT based approaches (i.e., Lord's χ^2 and likelihood ratio test) and the Mantel-Haenzel method detected 40% of the items to have DIF. Interestingly, several observed group mean differences were no longer significant when the DIF items were removed from the scale, which indicates that the observed mean differences were attributable to measurement artifacts rather than true mean differences between groups. Taking both languages and cultures into consideration, Ellis and Mead (2000) assessed the ME of the Sixteen Personality Factor (16PF) Questionnaire, which was administered in English to English-speaking Anglo-Americans and English-dominant Hispanic Americans, and in Spanish to Spanish-dominant Hispanic Americans and Spanish-speaking Mexicans. The authors hypothesized that the extent to which items have DIF is dependent on the extent to which language and/or culture differed in the two groups, so the least amount of DIF was expected in the English-speaking Anglo versus English dominant Hispanic Americans comparison (with no translation involved) while greater DIF was expected in the Anglo versus Spanish-speaking Mexicans comparison (due to both translation error and cultural differences). Results from the DFIT method showed that the pattern of DIF did not differ significantly across the three comparisons as hypothesized, although partially in the hypothesized direction.

Another line of research has examined ME of personality inventories using the MACS approach. Ghorpade, Hattrup and Lackritz (1999) explored the metric invariance of Locus of

Control, Self-Esteem and Higher Order Need Strength across US and Indian men and women. Equivalence of factor loadings was found across US and Indian men and women for Locus of Control; while equivalence of factor loadings of Self-Esteem and Higher Order Need Strength was found across US men, US women and Indian men but not Indian women. Nye, Roberts, Saucier and Zhou (2008) assessed ME of the Big Five Mini-Markers (Saucier, 1994) among Chinese, Greek, and American respondents and found configural invariance but not metric or scalar invariance. When the DIF items were removed from the scale, the group mean differences also changed and some of the conclusions were even reversed. For example, when the full set of items was examined, Greek respondents scored significantly higher on Conscientiousness than the Chinese who in turn scored higher than the Americans. However, when the scores were recalculated with the unbiased items, Americans showed higher Conscientious scores than Chinese. Again, these results show the importance of ME without which the true mean differences could be confounded with item bias and result in erroneous conclusions if we only look at observed mean differences.

All the cross-cultural studies on personality inventories discussed above investigated the effect of language and/or cultural differences on ME. These studies involved translation procedures whenever the respondents' native language was different from the language of the original scales. In the current study, we removed the effect of translation by examining ME across cultures with the items of the personality inventory written in a single language, English. If DIF is detected, we can conclude that it is not due to translation. We used both IRT and MACS approaches to examine ME of the Extraversion and Conscientiousness scales from the English version of Big Five Personality Inventory (John et al., 2008) across respondents from four countries: the United States, China, Australia and Spain. We developed our hypotheses

based on the culture proximity articulated by Schwartz (1999). That is, the United States and Australia belong to the English-speaking group and have much similarity in language and culture. Therefore, respondents from these two countries should share a more similar interpretation of the personality scales and the psychological constructs they measure. Spain belongs to the West Europe group and its native language is Spanish. But given the shared history and the similarity in these two languages, we expected to find some ME between the US and the Spanish groups. Finally, China belongs to the Far East group and the Chinese language is very different from English in terms of vocabulary and grammar. As Chinese share the least in common with Americans, we hypothesized that we would find the least ME and the largest mean difference in latent constructs.

METHOD

Sample

Respondents were anonymous individuals who took the online version of Big Five Personality Inventory test from www.outofservice.com. The only demographic information relevant to our study is nationality. In the original dataset the sample sizes of US and Australian respondents were much larger than those of Chinese and Spanish respondents. To make the sample sizes comparable across the four groups we randomly selected 40% of the US respondents and 4% of the Australian respondents. The respondents included in the subsequent analyses were 452 from the US, 277 from China, 489 from Australia and 424 from Spain.

Measures

The online personality inventory analyzed in this study is the 44-item BFI (John et al., 2008), which uses short phrases to represent the prototypical markers of Extraversion, Conscientiousness, Neuroticism, Agreeableness and Openness. The coefficient alpha averaged across the five subscales is .83 (John & Srivastava, 1999). For each item there are five response categories from *strongly disagree to strongly agree*. In the current study we analyzed two scales: Extraversion and Conscientiousness. The Extraversion scale has 8 items with a coefficient alpha of .88. Sample items included "I see Myself as Someone Who…" "Is talkative" and "Is reserved" (reverse coded). The Conscientiousness scale has 9 items with a coefficient alpha of .82. Sample items included "I see Myself as Someone Who…" "Does a thorough job" and "Can be somewhat careless" (reverse coded).

Analyses

Preliminary analyses. Before we ran the IRT and MACS analyses, we first reviewed the data and confirmed that responses to some of the items had been reverse coded as needed. To detect the respondents who chose the same option for all items in the scale without considering the reverse coded items, we made a scatter plot with the aggregated score of the positively worded items as the x-axis and the aggregated score of the negatively worded item as the y-axis. Data points from respondents who carefully read the items should lie closely to the diagonal line. Based on the scatter plot, we detected respondents whose data points lay at the lower right corner or the upper left corner of the coordinate. We removed the data of these respondents from subsequent analyses because these respondents may have chosen options blindly without carefully reading the items. Excluding these respondents should reduce problems associated with random or careless responses.

IRT approach. To meet the assumption of IRT models, we first checked the unidimensionality of the data. There are several methods of evaluating dimensionality and in the current study we used exploratory and confirmatory factor analysis (CFA) approaches. For the exploratory approach, we ran a principal component analysis (PCA) using SPSS 17.0 and examined the scree plot and factor loadings. If the scree plot shows a strong first factor and the items all have moderate to high loadings on the first factor, we can conclude that the data are sufficiently unidimensional for IRT analyses. As a supplement to the exploratory approach, we also ran a CFA analysis where we fit a one-factor model using LISREL v. 8.72 (Jöreskog & Sörborn, 2004). Then we examined fit indices such as the comparative fit index (CFI), the Tucker-Lewis index (which is sometimes called the non-normed fit index or NNFI), the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMSR). We compared the fit index with the recommended cutoff values: CFI $\geq$ .95, NNFI

≥ .95, RMSEA ≤ .06 and SRMSR ≤ .08 (Hu & Bentler, 1999). If the values of the fit index fall within the range of these cutoffs, then a unidimensional model fits the data well.

After we confirmed the unidimensionality of the data, we proceeded to run the IRT analyses. We used MULTILOG 7.03 (Thissen, Chen, & Bock, 2003) to estimate the discrimination and threshold parameters for each item for the reference group and the focal group. In the current study, we used the US sample as the reference group and the samples from the other three countries as three focal groups. We ran Baker's (1995) EQUATE 2.1 to put the item parameters from the reference group and the focal group on a common scale and obtained the NCDIF index for each item using the DFIT software (Raju, 1999). We detected the DIF items by comparing the NCDIF index of each item with the cutoff score. Traditionally the DFIT program suggests a critical value of .096 for items with five response options (Raju, 1999). However, empirical research suggests a much smaller cutoff to obtain more accurate results. In the current analyses we used the regression-based cutoff value, which is a composite of sample size, scale length, number of response options and their interactions (Tay, Nye, & Drasgow, 2010). Results from Tay et al.'s simulation studies have shown that the regression-based cutoff values performed better than the empirically derived cutoff values.

For the first round of the linking process, we used all the items of a scale as anchor items. Items with index values larger than the cutoff value were flagged as having DIF. For the second round, we used only the non-DIF items identified in the first round to do the linking and identified DIF items in the same way (Candell & Drasgow, 1988). As two rounds of linking are usually enough to identify the DIF items, we finalized the set of DIF items between the reference and each focal group after the two-step iterative linking. Those items with the least level of DIF

were used as referents in the subsequent MACS analysis. After we found the DIF items between the reference and the focal group, we constrained the parameters of non-DIF items to be equal and allowed the parameters of DIF items to vary freely, which makes the latent traits comparable within each pair. Then we examined the mean differences between the reference group and focal groups by examining the MULTILOG output from the concurrent estimation. As MULTILOG sets the focal group mean to zero, we can interpret the trait value of the reference group as the true mean difference between the reference and the focal group.

MACS approach. Before we conducted the CFA we dichotomized the data for diagonally weighted least squares (DWLS) estimation. DWLS estimation corrects for non-normality and is suitable for data measured on ordered categorical scales (e.g., Finney & DiStefano, 2006). We used DWLS for analyzing item-level categorical data, which violate the assumption of multivariate normality of the observed variables that is needed for maximum likelihood estimation.

As we discussed in the IRT approach, we examined the unidimensionality of the responses for each country. If the unidimensional model fits the data, configural invariance is established. To establish metric and scalar invariance, we constrained factor loadings and intercepts to be equal within each pair of reference and focal groups and compared the constrained model with the baseline model. In a baseline model parameters for all items were allowed to vary freely except the referent. The loading of this item was fixed at 1 in both the reference group and the focal group and its intercept was constrained to be equal across groups. In a series of constrained models, the loading and intercept for each individual item were constrained to be equal across groups in addition to the referent item. The baseline model and

constrained models were then compared. As the constrained model was nested in the baseline model, the observed change in χ^2 can be compared to the χ^2 distribution with 2 *df*. We used a Bonferroni correction to control for Type I errors. A significant χ^2 change indicates that the constrained parameters are not equivalent between groups and the studied item was flagged as having DIF. We did this for each item, one at a time to identify the DIF items, and then we constrained all the non-DIF items simultaneously to estimate the latent mean difference. As the latent mean is fixed to 0 for the reference group, the mean difference between the reference group and the focal group equals the latent mean of the focal group from the LISREL output.

Effect size analysis. The effect size of measurement nonequivalence in the IRT methodology was calculated using the Matlab script developed by Nye (2011). Input files included the item parameter and ability score files of the reference group, and the equated item parameter and the ability score files of the focal group. For the MACS approach the effect size was calculated using the program developed by Nye and Drasgow (2011). For both the reference and focal groups we input parameter files (e.g., factor loadings, theta-delta matrix, etc.) which were given in the LISREL output.

RESULTS

Extraversion

We first removed the data from those respondents who appeared to ignore the reverse coded items. The resulting sample sizes for each country were N = 452 for Australia, N = 269 for China, N = 389 for Spain and N = 481 for the United States. Table 1 shows the means and standard deviations of the 8 items for these four countries.

The PCA results supported unidimensionality for the data from Australia, Spain and the US. The first eigenvalue explained about 50% of the variance and all the items had moderate to high loadings (ranging from .50 to .83) on the first factor. For China, PCA suggested two components with eigenvalues larger than 1.0 which explained 36.69% and 13.78% of the variance respectively.

To further explore the dimensionality of the data, we ran CFA using DWLS analysis. We dichotomized the data based on the frequency of responses to each category in order to have about 50% of the respondents in each of the 2 dichotomized categories. Then we fit the one-factor model to the dichotomized data. For all the four countries, CFI, NNFI, RMSEA, and SRMR indices showed acceptable model fit for item-level categorical data (see Table 2). This suggests the second factor for the Chinese data was an artifact of the dichotomous data, which is called a difficulty factor (see Hulin et al., 1983, Chap. 8).

After we confirmed that the data were sufficiently unidimensional for each country, we proceeded to run IRT analyses to obtain estimated item parameters and respondents' latent traits. Then we used iterative linking to transform item parameters of the 3 focus groups (i.e., Australia,

China and Spain) to the reference group (i.e., the U.S.) and identified DIF items based on the composite cutoff score of scale length (i.e., 8 for Extraversion), sample size (i.e., the mean of the reference group and the focus group), number of options (n = 5 for BFI), and their interactions. For instance, we used all the 8 items in the scale as anchor items for the first round of linking between the US and China. Five items (i.e., item 1, 3, 4, 6, and 7) had NCDIF larger than the cutoff and thus were temporarily identified as DIF items. For the second round, we used only the 3 non-DIF items as the anchor items to do the linking. Four items, items 1, 3, 4, and 7 were identified as having DIF. As two rounds of linking are usually enough to identify the DIF items, we considered this set of 4 items as the final output of DIF items for the US-China comparison. For the U.S. and Australia comparison, the same set of DIF items (i.e., item 2 and 7) were identified for the two rounds. The same set of DIF items, items 2, 3, 5, 6, and 7 were flagged as DIF items between the US and Spain after two rounds of linking (see Table 3).

After we detected the DIF items, we did concurrent estimation for each reference group and focal group pair. Using the nonequivalence groups with anchor test (NEAT; Kolen & Brennan, 2004) framework, we considered the non-DIF items (item 2, 5, 6, and 8) as "common items" that both Americans and Chinese answered. We considered DIF items as "unique items" to each form of the scale. This method yielded a new scale of 12 items, consisting of 4 common items, 4 items unique to American respondents, and 4 items unique to Chinese respondents. Thus the scales for the reference group and the focal group were linked through the common items and the estimated item parameters and respondents' latent traits were comparable to each other.

Using the new set of estimated item parameters, the mean latent trait for the US was fixed at 0 by MULTILOG and the mean latent trait for China to be – 0.37. Therefore, the mean

difference of Extraversion between the US and Chinese respondents was 0.37, which indicated the Americans are moderately more extraverted than the Chinese. We used the same procedures for the other two pairs of the reference and the focal group but found almost no difference in Extraversion for the US-Australia comparison and the US-Spain comparison.

For the MACS approach, configural invariance was established as we confirmed the unidimensionality of the data for each country. Then we proceeded to examine metric and scalar invariance simultaneously. In the first step, the loading and intercept of each item were constrained to be equal, one item at a time. We evaluated χ^2 change in the fit statistic between the baseline model, which only constrained the referent item, and each constrained model against a cutoff value with Bonferroni correction for multiple comparisons. Those χ^2 changes larger than the cutoffs suggested model misfit and nonequivalence. As Table 4 shows, configural and scalar invariance was not established for one item (i.e., a DIF item) for the US-Australia comparison. Two items failed to achieve ME for the US-China comparison and the US-Spain comparison. Then we estimated latent mean differences between the reference and the focal group by constraining loadings and intercepts of *all* non-DIF items to be equal and left those of the DIF items to vary freely. For the US-China comparison, we constrained the loadings and intercepts of all the items except item 3 and 4 and obtained a latent mean difference of -0.5. For the US-Australia and the US-Spain comparisons, latent mean differences were -0.03 and -0.01 respectively. Thus, these results are quite consistent with the IRT results.

Conscientiousness

As we did for the Extraversion scale, we removed the data from those who seemed to ignore the reverse coded items. Sample sizes for each country were N = 453 for Australia, N =

254 for China, N = 391 for Spain, and N = 402 for the United States. Table 5 shows the means and standard deviations of the 8 items for the four countries. The PCA results suggested 2 factors for the 9-item scale for the US respondents. When we fit the one-factor model to the US data, the loadings estimated by LISREL were quite high for Item 1 through 8 but extremely low for Item 9. To maintain a unidimensional structure for the reference group, we deleted Item 9 and used an 8-item scale for subsequent analyses. The PCA results supported unidimensionality of the reduced scale for all the four countries. The first eigenvalue explained about 40% - 49% of the variance and all the items had moderate to high loadings (ranging from .51 to .80) on the first factor. Table 4 shows the indices for the unidimensional model fit by DWLS. For all the four countries, CFI, NNFI, RMSEA, and SRMR indices showed acceptable model fit for item-level categorical data (see Table 6).

Following the same procedures as we did for the Extraversion scale, we performed linking and identified DIF items between the reference group and each focal group. Table 7 shows the DIF items identified by DFIT after two rounds of linking. Four items were flagged as DIF item for the US-Australia comparison and the US-Spain comparison. Five items were found to have DIF between the US and China. We used the non-DIF items identified at the second round of linking as the common items for the concurrent estimation. The mean latent trait was fixed at 0 for the US, – 0.31 for China, 0.05 for Australia and -0.19 for Spain. Therefore, the mean difference of Conscientiousness between the US and Chinese respondents was moderately large and the one between the US and Spanish respondents was small in magnitude. Almost no difference was found for the US-Australia comparison.

For the MACS approach, configural invariance was established as we confirmed the unidimensionality of the 8-item scale for each country. As Table 8 showed, 2 items failed to achieve ME for the comparisons between the US and each of the three other countries. We constrained the loadings and intercepts of all non-DIF items and obtained the latent mean difference of -0.16 for the US-China comparison, -0.02 for the US-Australia and -0.12 for the US-Spain comparisons. These results suggest a smaller difference for the US-China comparison than found in the IRT analyses, but highly consistent findings for the other two comparisons.

Effect size analyses. Table 9 shows the effect sizes for measurement nonequivalence of Extraversion items between each pair of the reference and focal groups using the IRT and the MACS approaches. Overall, the findings by the IRT and MACS approaches were highly consistent in regard to the magnitude of the effect size. That is, the items with large effect sizes in the IRT methodology also tended to have large effect sizes for the MACS approach. For the MACS approach, the referent item was set to have an effect size of zero. The IRT approach found the referent item to have the smallest effect size. The largest effect size was found for Item 2 between the U.S. and Spain. Effect sizes of similar magnitudes were also found for Items 3 and 4 between the US and China; and for Item 2 between the US and Australia. None of the effect sizes exceeded the medium magnitude of .5 defined in Cohen's terminology of group mean difference (Cohen, 1992).

For the Conscientiousness scale, the referent items again showed the smallest effect size (see Table 10). The largest effect sizes were found for Items 2 and 4 between the US and China; and for Item 1 between the US and Spain. Although the magnitude of most effect sizes found by the IRT and the MACS approaches showed a consistent pattern, there were some discrepancies

in the estimates between these two approaches such as the effect size of Item 1 between the US and Australia and Item 5 between the US and China. Again no items had an effect size larger than a medium magnitude.

DISCUSSION

There has been a plethora of cross-cultural studies on personality. However, it is important that ME is established before meaningful group comparisons are made. This paper examined the ME of two scales from the BFI across 4 countries. Different from most cross-cultural personality research, the personality inventory in this study was administered only in English and thus removed the effect of translation. Using the IRT approach to examine the ME of the Extraversion scale, we found 2 DIF items (Item 1 & 7) for the US-Australia comparison, 4 DIF items (Item 1, 3,4 & 7) for the US-China comparison and 5 DIF items (Item 2, 3, 5, 6, & 7) for the US-Spain comparison. Meanwhile, the MACS approach detected half the amount of DIF items: 1 item (Item 2) for the US-Australia comparison and 2 for the US-China (Item 3 & 4) and the US-Spain (Item 2 & 6) comparisons. Note that the DIF items detected by the MACS approach were those with the highest NCDIF values of the IRT approach. Therefore, it appears that the MACS approach has a lower Type I error rate but also less power due to the Bonferroni correction. The results for the Conscientious scale were similar. The IRT approach detected 4 DIF items (Item 1, 2, 4, & 5) for the US-Australia comparison, 5 DIF items for the US-China comparison (Item 1, 2, 3, 4, & 5) and 4 DIF items for the US-Spain comparison (Item 1, 3, 5, & 7). Using the MACS approach, we found only 1 DIF item for each of the comparisons. Therefore, for both scales, there were DIF items for all the comparisons between the reference group and the focal group. The numbers of DIF items were similar across comparisons and thus were not exactly consistent with the pattern we hypothesized based on cultural proximity.

To examine the mean differences between the reference group and each focal group, we circumvented the effects of DIF items by viewing the scales from the perspective of the NEAT

design (Kolen & Brennan, 2004). This allowed us to retain all the items for our analyses but not confound results because of DIF. The largest mean differences in Extraversion and Conscientiousness were found between the US and Chinese respondents by both the IRT and MACS approaches. This result was consistent with previous findings that Chinese respondents scored lower than the US respondents in Extraversion and Conscientiousness (McCrae, Costa & Yik, 1996; Nye et al., 2008; Yang, 1986). Little difference was found in Extraversion and Conscientiousness between the US and the Australian respondents. For the US-Spain comparison, we found no difference in Extraversion and a small effect in Conscientiousness. Therefore, the pattern of the latent mean differences was consistent with our hypotheses. The largest differences were found between the US and China, which have the most dissimilarity in language and culture. The US-Australia showed the least differences due to their shared culture and language.

Finally, our study examined the magnitude of measurement nonequivalence by calculating the effect size in both the IRT and the MACS methodologies. Overall, these two approaches found consistent patterns of effect size for each item between the reference and focal groups. In addition, those items with the largest effect sizes were also those with the largest χ^2 in the significance test. Although both the effect size and the significance test indicate the "amount" of measurement nonequivalence, they are fundamentally different in terms of interpretation. For the significance test, the χ^2 indexes the *probability* of measurement nonequivalence while the effect size provides a standardized metric of *magnitude* that is comparable to other effect size measures. Most important of all, none of the DIF items were found to have more than a medium effect. That is, although almost half of the items of each scale were flagged as DIF items in the significance test, their practical importance may not be as influential as the results of the significance test implied.

Implications. We believe our study has important implications for future research. First, our study gave an example of examining ME for cross-cultural studies using both the IRT and the MACS approaches. We elaborated the procedures developed in the most recent ME research and test equating literature. For example, in our DIF study we used the NEAT design to treat non-DIF items as common items for the reference and focal groups while we treated the DIF items as unique items. In this way we did not have to remove all the DIF items and thus retained as much information as possible from the personality inventory. In addition, we compared the χ^2 of the NEAT design with the sum of χ^2 of the separate estimation for the reference and focal groups. The χ^2 change was not significant, which indicates that the NEAT design fit as well as the separate estimation while equating the item parameters and ability scores for the reference and focal groups. Therefore, the NEAT design was an effective method for including both DIF and non-DIF items in the estimation for the reference and focal groups without confounding the results with DIF.

Second, our study to some extent confirmed that there was more similarity in the meaning of Extraversion and Conscientiousness to respondents from proximate cultures than respondents from distant cultures. While the numbers of DIF items found between the US and three other countries were not exactly consistent with the pattern we hypothesized, the mean differences were in line with our hypotheses. That is, the differences in both Extraversion and Conscientiousness between the US and China were found to be the largest while the differences between the US and the other two Western or English-speaking countries were either much smaller or almost negligible. Although we had only four countries in our study, these results were consistent with previous findings that proximate cultures had similar personality profiles

and that there was a clear contrast of American and European cultures with African and Asian cultures (Allik & McCrae, 2004).

Finally, in our study the BFI was administered only in English and thus controlled for the effect of language. Although the IRT approach found half of each 8-item scale to be DIF items and the MACS approach found at least one DIF item between the reference and focal groups, the results of our effect size analyses suggested that none of these DIF items had larger than a medium effect size. Therefore, we believe that effect size analyses should be conducted in addition to significance test to give researchers a better idea of the practical significance of measurement nonequivalence.

From a practical perspective, our results showed promise for administering personality inventories in English to non-native speakers with English proficiency. Given our findings that the magnitude of measurement nonequivalence was not large and the latent mean differences were consistent with results from previous studies in which the personality inventories were translated into respondents' native languages, it appears that personality inventories can accurately reflect the respondents' standing on the latent traits even when they respond in their second language. If this result is confirmed by future studies, we will have more confidence in administering personality inventory in English (or some common language) to all proficient bilinguals without devoting time and effort to translation.

Limitations. Like most cross-cultural research, we delimited culture by countries and assumed that the within-culture variance should be smaller than between-culture variance. However, in countries like the US and China, which have diverse cultures in different regions, we should generalize our results with caution. Perhaps those competent English-speakers in

China were from more developed regions and thus were more influenced by Western values and shared a more similar culture with Americans than traditional Chinese. If we had a representative sample of Chinese respondents, we might expect even larger group mean differences in personality traits between Americans and Chinese.

Also, although we assumed that all our respondents were competent English-speakers, language proficiency may still affect ME because language itself may activate some personality traits among bilinguals (Ramírez-Esparza, Gosling, Benet-Martínez, Potter, & Pennebaker, 2006). Future research should focus on how language similarity may influence personality traits measured by inventories written in respondents' native language or in a proficient second-language.

CONCLUSION

The IRT and the MACS approaches found similar numbers of items with measurement nonequivalence of Extraversion and Conscientious Scale for the US-Australia, US-China and US-Spain comparisons. The IRT approach tended to detect more items with measurement nonequivalence than the MACS approach. After removing the measurement artifact we found mean differences in Extraversion and Conscientious between the US and three other countries to be consistent with our hypothesized cultural similarity. The effect size analyses showed that none of the DIF items had large effect. Therefore, administering personality inventories in English to participants with English proficiency, but from other cultures and speaking different languages, can be a convenient alternative to translation and thus save time and money.

TABLES

Table 1. Mean and SD of the Extraversion items of BFI for the four countries.

	The US N = 446		Australia N = 452		China N = 269		Spain N = 389	
Items	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1. Is talkative	3.64	1.24	3.71	1.25	3.28	1.34	3.60	1.25
2. Is reserved (R)	2.91	1.25	2.88	1.29	2.81	1.20	2.52	1.34
3. Is full of energy	3.53	1.21	3.43	1.17	3.66	1.11	3.69	1.10
4. Generates a lot of enthusiasm	3.56	1.16	3.59	1.05	3.67	1.09	3.50	1.14
5. Tends to be quiet (R)	2.99	1.42	2.94	1.36	2.51	1.30	2.98	1.41
6. Has an assertive personality	3.52	1.19	3.38	1.21	3.24	1.17	3.49	1.03
7. Is sometimes shy, inhibited (R)	2.70	1.31	2.52	1.28	2.21	1.07	2.44	1.37
8. Is outgoing, sociable	3.54	1.31	3.60	1.24	3.17	1.24	3.62	1.26
Average of Means and SDs	3.30	1.26	3.26	1.23	3.07	1.19	3.23	1.24

Table 2. Unidimensionality fit indices for the 8-item Extraversion scale of BFI.

	CFI	NNFI	RMSEA	SRMR
United States (N = 446)	0.98	0.98	0.083	0.062
Australia (N = 452)	0.98	0.97	0.082	0.078
China (N = 269)	0.98	0.97	0.063	0.087
Spain (N = 389)	0.98	0.98	0.070	0.069

Table 3. NCDIF indices from iterative linking for the Extraversion scale.

	Round 1 of linking (using all items as linking items)			Round 2 of linking (using only non-DIF items)		
Items	Australia	China	Spain	Australia	China	Spain
1. Is talkative	0.014	0.041*	0.001	0.009	0.037*	0
2. Is reserved (R)	0.099*	0.010	0.198*	0.144*	0.012	0.211*
3. Is full of energy	0.004	0.103*	0.037*	0.004	0.122*	0.029*
4. Generates a lot of enthusiasm	0.021	0.087*	0.002	0.009	0.107*	0.003
5. Tends to be quiet (R)	0	0.006	0.026*	0.003	0.001	0.027*
6. Has an assertive personality	0.008	0.036*	0.053*	0.013	0.029	0.060*
7. Is sometimes shy, inhibited (R)	0.027*	0.059*	0.044*	0.036*	0.041*	0.052*
8. Is outgoing, sociable	0.020	0.029	0.010	0.006	0.023	0.005
Cutoff	0.024	0.031	0.026	0.024	0.031	0.026

Notes. Items with * were those with NCDIF values larger than the cutoff scores and were thus flagged as DIF items. Cutoff scores were set based on the method proposed by Tay et al. (2010).

Table 4. χ^2 change in fit statistics between the constrained model and the baseline model for the Extraversion scale.

Items	U.S. vs. Australia	U.S. vs. China	U.S. vs. Spain
1. Is talkative	1.85	9.04	Referent
2. Is reserved (R)	17.22*	2.09	17.85*
3. Is full of energy	1.87	13.43*	5.5
4. Generates a lot of enthusiasm	6.99	15.09*	0.2
5. Tends to be quiet (R)	Referent	Referent	1.27
6. Has an assertive personality	2.38	1.67	14.86*
7. Is sometimes shy, inhibited (R)	3.97	5.43	5.73
8. Is outgoing, sociable	3.42	1.97	1.6

Notes. The cutoff was $\chi^2(2) = 13.1426$ after Bonferroni correction for 7 comparisons ($\alpha = .01/7 = 0.0014$).

Table 5. Mean and SD of the Conscientiousness items of BFI for the four countries

	The US N = 402		Australia N = 453		China N = 254		Spain N = 391	
Items	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1. Does a thorough job	3.89	1.07	3.79	1.06	3.53	1.20	3.72	1.14
2. Can be somewhat careless (R)	3.02	1.24	2.83	1.22	2.31	1.11	2.73	1.26
3. Is a reliable worker	4.14	1.01	4.20	0.97	3.97	1.14	3.95	1.10
4. Tends to be disorganized (R)	2.96	1.41	2.85	1.41	3.09	1.30	2.69	1.40
5. Tends to be lazy (R)	2.99	1.30	2.74	1.30	2.85	1.37	2.82	1.40
6. Perseveres until the task is finished	3.65	1.16	3.59	1.14	3.59	1.13	3.60	1.23
7. Does things efficiently	3.77	1.00	3.81	0.98	3.57	1.21	3.88	1.03
8. Make plans and follows through with them	3.43	1.05	3.52	1.13	3.11	1.21	3.29	1.21
Average of Means and SDs	3.48	1.16	3.42	1.15	3.25	1.21	3.34	1.22

Table 6. Unidimensionality fit indices for the 8-item Conscientiousness scale.

	CFI	NNFI	RMSEA	SRMR
United States (N = 402)	0.98	0.97	0.077	0.071
Australia (N = 453)	1.00	1.00	0.027	0.042
China (N = 254)	0.99	0.98	0.050	0.069
Spain (N = 391)	0.98	0.97	0.063	0.070

Table 7. NCDIF indices from iterative linking for the Conscientiousness scale.

	Round 1 of linking (using all items as linking items)			Round 2 of linking (using only non-DIF items)		
Items	Australia	China	Spain	Australia	China	Spain
1. Does a thorough job	0.064*	0.085*	0.081*	0.049*	0.038*	0.092*
2. Can be somewhat careless (R)	0.022	0.215*	0.012	0.029*	0.126*	0.008
3. Is a reliable worker	0.022	0.046	0.040*	0.014	0.043*	0.045*
4. Tends to be disorganized (R)	0.057*	0.311*	0.032*	0.080*	0.552*	0.022
5. Tends to be lazy (R)	0.058*	0.176*	0.039*	0.068*	0.354*	0.031*
6. Perseveres until the task is finished	0.021	0.035*	0.001	0.017	0.032	0
7. Does things efficiently	0.014	0.013	0.022	0.008	0.001	0.033*
8. Make plans and follows through with them	0.003	0.061*	0.020	0.007	0.024	0.013
Cutoff	0.025	0.034	0.027	0.025	0.034	0.027

Notes. Items with * were those with NCDIF values larger than the cutoff scores and were thus flagged as DIF items. Cutoff scores were set based on the method proposed by Tay et al. (2010).

Table 8. χ^2 change in fit statistics between the constrained model and the baseline model for the Conscientiousness scale.

Items	U.S. vs. Australia	U.S. vs. China	U.S. vs. Spain
1. Does a thorough job	16.62*	4.38	13.33*
2. Can be somewhat careless (R)	6.58	20.37*	6.72
3. Is a reliable worker	8.51	2.94	5.65
4. Tends to be disorganized (R)	2.78	6.6	2.47
5. Tends to be lazy (R)	7.78	1.23	0.86
6. Perseveres until the task is finished	4.72	7.58	Referent
7. Does things efficiently	7.15	Referent	7.28
8. Make plans and follows through with them	Referent	2.24	3.27

Notes. The cutoff was $\chi^2(2) = 13.1426$ after Bonferroni correction for 7 comparisons ($\alpha = .01/7 = 0.0014$).

Table 9. Effect sizes of measurement nonequivalence for Extraversion.

	U.S. vs. Australia		U.S. vs. China		U.S. vs. Spain	
Items	IRT	MACS	IRT	MACS	IRT	MACS
1. Is talkative	0.100	0.076	0.182	0.282	0.016	0
2. Is reserved (R)	0.334	0.276	0.115	0.133	0.433	0.391
3. Is full of energy	0.061	0.085	0.357	0.320	0.185	0.188
4. Generates a lot of enthusiasm	0.095	0.160	0.363	0.329	0.057	0.018
5. Tends to be quiet (R)	0.051	0	0.030	0	0.145	0.112
6. Has an assertive personality	0.114	0.106	0.128	0.119	0.234	0.294
7. Is sometimes shy, inhibited (R)	0.184	0.147	0.212	0.194	0.214	0.231
8. Is outgoing, sociable	0.077	0.120	0.163	0.135	0.070	0.102

Table 10. Effect sizes of measurement nonequivalence for Conscientiousness.

	U.S. vs. Australia		U.S. vs. China		U.S. vs. Spain	
Items	IRT	MACS	IRT	MACS	IRT	MACS
1. Does a thorough job	0.235	0.503	0.202	0.282	0.327	0.387
2. Can be somewhat careless (R)	0.172	0.265	0.431	0.484	0.089	0.258
3. Is a reliable worker	0.152	0.333	0.295	0.218	0.239	0.213
4. Tends to be disorganized (R)	0.260	0.199	0.583	0.315	0.142	0.178
5. Tends to be lazy (R)	0.259	0.292	0.420	0.158	0.155	0.082
6. Perseveres until the task is finished	0.128	0.277	0.220	0.359	0.016	0
7. Does things efficiently	0.094	0.309	0.038	0	0.228	0.258
8. Make plans and follows through with them	0.085	0	0.206	0.172	0.134	0.177

REFERENCES

Allik, J., & McCrae, R. R. (2004). Toward a geography of personality traits: Pattern of profiles across 36 cultures. *Journal of Cross-Cultural Psychology, 35(1),* 13-28.

Baker, F. (1995). EQUATE (Version 2.1): Computer program for equating two metrics in item response theory [Computer software]. Madison, WI: University of Wisconsin, Laboratory of Experimental Design.

Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods, 6(1),* 15-43.

Berry, J. W. (1969). On cross-cultural comparability. *International Journal of Psychology, 4(2),* 119-128.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19(4),* 309-321.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12(3),* 253-260.

Candell, G. L., & Hulin, C. L. (1986). Cross-language and cross-cultural comparisons in scale translations: independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 17(4),* 417-440.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hilldale, NJ: Earlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1),* 155-159.

Cole, M. S., Bedeian, A. G., & Feild, H. S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods, 9(3),* 339-368.

Donovan, M. A., Drasgow, F., & Prost, T. M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology, 85(2),* 305-313.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70(4),* 662-680.

Ellis, B. B. (1995). A partial test of Hulin's psychometric theory of measurement equivalence in translated tests. *European Journal of Psychological Assessment, 11(3),* 184-193.

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(2),* 133-148.

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(5),* 787-807.

Finney, S. J., & DiStefano, C. (2006). *Non-normal and categorical data in structural equation modeling: A second course* (pp. 269-314). Greenwich, CT: Information Age Publishing.

Ghorpade, J., Hattrup, K., & Lackritz, J. R. (1999). The use of personality measures in cross-cultural research: A test of three personality scales across two countries. *Journal of Applied Psychology, 84(5),* 670-679.

Hannum, K. M. (2007). Measurement equivalence of 360°- assessment data: Are different raters rating the same constructs? *International Journal of Selection and Assessment, 15(3),* 293-301.

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.

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(2),* 192-218.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of Cross-Cultural Psychology, 18(2),* 115-142.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspective. 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. G., & Sörborn, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software.

Jöreskog, K.G., & Sörbom, D. (2004). LISREL 8.7 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

King, W., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80(6),* 643-651.

Kirk, R. E. (2006). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference*. DOI: 10.1016/j.jspi.2006.09.011.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83(5),* 693-702.

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

McCrae, R. R., Costa, P. T., & 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). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89(3),* 407-425.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing, 5(3),* 279-300.

Meriac, J. P., Woehr, D. J., & Banister, C. (2010). Generational differences in work ethic: An examination of measurement equivalence across three cohorts. *Journal of Business Psychology, 25,* 315-324.

Mueller, K., Liebig, C., & Hattrup, K. (2007). Computerizing organizational attitude surveys: An investigation of the measurement equivalence of a multifaceted job satisfaction measure. *Educational and psychological measurement, 67(4),* 658-678.

Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies, 26,* 573-596.

Nye, C. D. (2011). *The development and evaluation of effect size measure for IRT and CFA studies of measurement equivalence.* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Champaign, IL.

Nye, C. D., & Drasgow, F. (2011, April 4). Effect size indices for analysis of measurement equivalence: Understanding the practical importance of difference between groups. *Journal of Applied Psychology*. Advanced online publication. doi: 10.1037/a0022955

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.

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: integrating correlational and experimental approaches. *Organizational Research Methods, 7(1)*, 27-65.

Raju, N. S. (1999). DFIT5P: A Fortran program for calculating DIF/DTF [Computer program]. Chicago, IL: Illinois Institute of Technology.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353-368.

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J. P., & Pennebaker, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of Research in Personality, 40,* 99-120.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20(3),* 643-671.

Ryan, A. M., West, B. J., & Carr, J. Z. (2003). Effects of the terrorist attacks of 9/11/01 on employee attitudes. *Journal of Applied Psychology, 88(4),* 647-659.

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assesses' perceptions and reactions. *International Journal of Selection and Assessment, 11,* 194-205.

Samejima, F. (1969). Estimation of latent ability using response pattern of graded scores. *Psychometrika Monograph, 17,* 1-100.

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

Saucier, G., & Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality, 69(6),* 847-880.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martinez, V. (2007). The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38(2),* 173-212.

Schwartz, S. H. (1999). A theory of cultural values and some implications for work. *Applied Psychology: An International Review, 48,* 23-47.

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*(6), 1292-1306.

Tay, L., Nye, C. D., & Drasgow, F. (2010). *Assessing DIF with DFIT: A regression-based approach for controlling type I error rates.* Manuscript submitted for publication.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of DIF using the parameters of items response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillside, NJ: Erlbaum.

Vandenberg, R. J., & Lance, C. E. (2004). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations, for organizational research. *Organizational Research Methods, 3(1),* 4-70.

Woehr, D. J., Sheehan, M. K., & Bennett, W. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology, 90(3),* 592-600.

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.