**Introduction**

Personality traits are important to the field of Industrial and Organizational Psychology in that they have been proved to predict a variety of work-related outcomes, including turnover (Salgado, 2000), task performance (Barrick & Mount, 1991; Salgado, 1997; Hrutz & Donovan, 2000; Hogan & Holland, 2003), organizational citizenship behavior (OCB; Borman, Penner, Allen, & Motowidlo, 2001), counterproductive work behavior (CWB; Donnellan, Spilman, Garcia, & Conger, 2014), leadership (Judge, Bono, Ilies, & Gerhardt, 2002), and job satisfaction (Judge, Heller, & Mount, 2002). In personnel selection, the good criterion-related validity along with their weak correlation with intelligence (Tett, Jackson, & Rothstein, 1991) have made personality tests an ideal supplement for intelligence tests.

However, comparisons across groups are meaningless if the test is lacking measurement equivalence (ME; Drasgow & Kanfer, 1985). Without ME, it’s hard to know if the observed mean score difference is due to true group difference or the relationships that are different across groups between the latent variable and the observed score (Raju, Laffitte, & Byrne, 2002). According to Drasgow (1984), ME was obtained when participants from different groups had the same expected observed score as long as they were at the same latent trait level. To test for ME in cross-cultural personality tests, we are not being paranoid, because over the years, measurement non-equivalence has been found in items on a variety of cross-cultural personality tests, including but not limited to the English-language version of the Trier Personality Inventory (TPI; Elllis, Becker, & Kimmel, 1993), the English-language version of the NEO Personality Inventory (NEO-PI; Huang, Church, & Katigbak, 1997), the Big Five Mini-Markers (Saucier, 1994; Nye, Roberts, Saucier, & Zhou, 2008), and the Rosenberg Self-esteem Scale (Baranik, Lakey, Lance, Hua, & Meade, 2008). The prevalence of measurement non-equivalence in personality tests makes it necessary to always assess ME before scores are compared across groups or any selection decisions are made based upon these scores.

The two major approaches to study ME are Confirmatory Factor Analytic (CFA) mean and covariance structure (MACS), and Differential Item Functioning (DIF). The former examines whether a common factor model exists across groups (Raju et al., 2002) and focuses on testing three types of measurement invariance, which are configural, metric, and scalar invariance (Vandenberg & Lance, 2000). According to Horn and McArdle (1992), configural invariance should be achieved before the other two types of measurement invariance can be teste. Configural invariance is the weakest type of ME that stands for the existence of the same number of factors and similar loading patterns across groups. Metric variance refers to factor loadings being invariant across groups. It is in the middle in terms of invariance strictness. Scalar invariance, the strongest form of invariance, implies that when items regress on latent variables, they have the same intercepts across groups (Steenkamp & Baumgartner, 1998).

The alternative approach to studying ME is the IRT-based differential item functioning (DIF). DIF is different from the CFA approach in several ways. First, CFA tests the three different degrees of ME one after one, while the IRT DIF tests the invariance of item discrimination (analogous to factor loadings in CFA) and location parameters (analogous to intercepts in CFA) at the same time. This is to say that under the DIF approach, metric and scalar invariance are tested simultaneously (Stark, Chernyshenko, & Drasgow, 2006a). Second, the nonlinear relationship posited by IRT between the latent construct and the true score at item/subscale level is equally tenable (when responses are polytomously scored) or even more appropriate (when responses are dichotomously scored) than the linear relationship assumed by the CFA approach (Raju et al., 2002). Third, the IRT context takes into consideration the compensatory nature of DIF (Raju, van der Linden, & Fleer, 1995; Raju et al., 2002), an issue that’s rarely seen discussed in the CFA context. Fourth, in IRT, besides item parameter estimates, we are also able to obtain the item characteristic curves (ICCs). These plots provide extra information, such as whether the DIF is uniform or non-uniform (Wang, Tay, & Drasgow, 2013), which indicates the source of DIF (LaPalme, Wang, Joseph, Saklofske, & Yan, 2016). Lastly, within the IRT framework, we can assess DIF using the ideal point model, which some previous studies believed were more appropriate for self-reported attitude and personality assessment (Chernyshenko, 2002; Stark, Chernyshenko, Drasgow, & Williams, 2006b). Therefore, in the current study, we examined ME via the IRT-based DIF approach.

Model selection is the first and probably the most important issue when adopting the IRT approach. The dominance model is widely accepted and used in IRT analysis. It derives from Likert’s (1932) approach to analyzing rating scales, and assumes that the higher a participant’s trait level, the morel likely she will answer positively. But the ideal point model should not be neglected. Drasgow, Chernyshenko, and Stark (2010) pointed out that the approach deriving from Thurstone (1928) was superior to the dominance model for personality assessment by successfully modelling intermediate item response and having better model-data fit. Also, as discussed above, the ideal point model was found to be more suitable if the trait assessment is self-reported (Tay & Drasgow, 2012). We are unable to find any cross-cultural personality DIF studies that have successfully compared empirically the performance of the two types of IRT models. LaPalme and colleagues (2016) had to drop the ideal point model from the DIF analysis for an emotional intelligence (EI) measure, and proceed with only the dominance model because of the severe misfit of GGUM. The bad fit, according to the authors, was probably due to the fact that the Wong and Law Emotional Intelligence Scale (WLEIS; Wong & Law, 2002) that they used was an ability measure rather than a trait measure. O’Brien and LaHuis (2011) examined DIF on personality tests under both the dominance and the ideal point model, but it was not in a cross-cultural setting. The comparison was between a group of applicants and a group of incumbents. In Carter, Dalal, Zickar, and Adams (2009), the DIF approach was applied under the the GGUM to examine the effects of vague quantifiers on making an item more, but no comparison was done between different IRT models.

Under both the dominance model and the ideal point model, DIF detection adopts the null hypothesis significance testing (NHST) paradigm, which provides information on only the existence but not the magnitude of DIF. Item selection decisions based solely on this paradigm could lead to deleting items with statistically significant yet trivial DIF that is barely meaningful. This is especially likely when the sample size is large. In order to have a more accurate understanding of the effects of DIF, we also adopted DIF effect size (Nye, 2011) in the current study.

In summary, the current study intended to examine measurement equivalence of some facet scales of the Comprehensive Personality Scale (CPS; Wang, 2013) via an IRT DIF method. The analysis will be done across the American and Chinese cultures. Samejima’s Graded Response (SGR; Samejima, 1969) model was applied in the dominance IRT model framework, while the Generalized Graded Unfolding Model (GGUM; Roberts, Donoghue, & Laughlin, 2000) was applied to represent the ideal point model. We examined model-data fit first under both models. Via NHST we assessed DIF existence with both models, and DIF effect sizes were computed to figure out DIF magnitude. Finally, we evaluated the existence and strength of intermediate items by observing and comparing item characteristics (ICC) and item paramters before and after the responses were dichotomized.

**The Comprehensive Personality Scale (CPS)**

The CPS is a result of years of work in Dr. Fritz Drasgow’s lab, and it was developed using the ideal point scale construction approach (Wang, 2013; Chernyshenko, Stark, Drasgow, & Roberts, 2007). The CPS consists of 440 items that cover a full set of 22 personality facets derived from the traditional Big-Five Personality Model. For example, the extraversion dimension was extended to the dominance, sociability, excitement, and energy facets. More than 100 items were originally written for each facet, and 20 of them were carefully selected to represent each facet. In terms of item extremity, each facet has approximately equal numbers of statement reflecting high, medium, and low trait levels (Wang, 2013).

***Measurement Equivalence of the CPS.*** Wang (2013) conducted DIF analysis for the complete CPS across two American groups (undergraduate students and MTurk workers). The analysis was carried out under the GGUM only. We haven’t found any studies investigating ME of the CPS in a cross-cultural setting, and comparing the performance of the dominance IRT model vs. the ideal point model. Therefore, in the current study, we assessed ME of two of the CPS scales across two cultures under two different IRT models.

**The IRT Approach to Measurement Equivalence**

Item response theory (IRT) is an alternative to classical test theory (CTT). Unlike CTT, whose analysis unit is the whole test (Hambleton, Swaminathan, & Rogers, 1991), IRT focuses on individual item responses and connecting them with the latent trait measured by the test (Drasgow & Hulin, 1990).

There are two major types of IRT models, one is the dominance model, and the other is the ideal point model. The 2-parameter logistic model (2PLM) and Samejima’s (1969) Graded Response Model (SGRM) are two representative dominance models to analyze dichotomous and polytomous personality measures, respectively. For the ideal point model, the General Graded Unfolding Model (GGUM; Roberts et al., 2000) has gained a lot of attention these years. The difference between the dominance model and the ideal point model lies in their assumptions about response processes. The dominance IRT methods, deriving from Likert’s 1932 rating scales development and analysis approach, assume that the higher the respondent’s trait level, the more likely she will answer positively (Drasgow et al., 2010). Whereas the ideal point methods, inspired by a series of Thurstone’s (1927, 1928, 1929) studies on measuring attitudes, hypothesize that the closer the statement is to a respondent’s trait level, the higher the probability of endorsement (Drasgow et al., 2010).

***The dominance IRT models: 2PLM and SGRM***

In personality tests, the 2-parameter logistic model (2PLM) and Samejima’s Graded Response model are two very widely used IRT models. The 2PLM is applicable to dichotomous responses, while SGRM deals with polytomous response data.

The item response function (IRF) for the 2PLM is:

*Pi (θ)* is the probability of a random respondent correctly answering Item *i* correctly. Since the responses are dichotomous, the probability of answering this item wrong then is given by:

There are two item parameters in a 2PLM. *ai*is the discrimination parameter that represents the degree to which an item separates latent trait levels that are close to each other (Maurer, Raju, & Collins, 1998). The larger *ai* is, the steeper the IRF. *bi*is the difficulty parameter. It is the point on the latent trait (*θ*) scale where the probability of a correct response is equal to 0.5. The larger the difficulty parameter, the harder the item. D is the scaling factor that lets the logistic function resemble as close as possible the normal ogive curve, and is usually set equal to 1.702 (Valbuena, 2003). Exp stands for an exponential function.

Samejima’s (1969) Graded Response Model is an enhancement of the 2PLM (Kosinski, 1999) and one of the most popular polytomous models in personality research. Under SGR, a polytomous response is broken down to a series of binary response sets by boundary response function (BRF), which is obtained by gradually merging response options (Kosinski, 1999). The probability of selecting response option *k* equals the probability of endorsing response option kand higher minus that of endorsing response option *k*+1 and higher. The probability of selecting option *k* on item *i* is given by:

*BRFi,k(θ)* is the boundary response function for option *k* on item *i*, and is calculated using 2PLM, so the item paramters (*ai*, *bi,k*)and scaling constant (*D*) used are the same as in 2PLM.

***The ideal point model: General Graded Unfolding Model (GGUM)***

The ideal point models are not as well developed as the dominance models, and among the few ideal point models, the most employed is the the General Graded Unfolding Model (GGUM; Roberts et al., 2000), which is applicable to both binary or graded responses. As discussed above, ideal point models assume a response process different from dominance models, and GGUM, according to Roberts et al. (2000), was developed based on four basic premise about the response process. The first premise is that an individual tends to agree with the item with trait level that’s close to her own trait level. The second premise is that a respondent disagrees with an item because the item trait level is either higher or lower than her own trait level. Similarly, a person closer to an item on the latent trait continuum can also agree with this item from either above or below. The third premise is that subjective responses (not observed responses) to attitude statements follow a cumulative item response model. The last premise is that an individual is equally likely to agree with an item located either *h* unites above or below her position on the attitude continuum. Developed from the four premises above, the formal definition of the GGUM is:

This function gives the probability associated with the *j*th respondent’s observable response to the *i*th item. *Zi* is the observable response to item *i*, and z ranges from 0 to C, with 0 standing for the strongest level of disagreement, and C standing for the strongest level of agreement. C equals the number of response options minus 1. M equals 2\*C+1, representing the number of subjective response categories minus 1. *αi* is the discrimination parameter, and *δi* is the location parameter of item i on the latent trait continuum. *τik* is the location of the *k*th subjective response category threshold on the theta continuum relative to the location of the *i*th item. The *τik*s are symmetric about the point (*θj* - *δi*) = 0.

***Data-model fit***

The dominance models are predominant in scale development and analysis, but generally work consistently well only in the context of cognitive ability testing (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001), because this is a domain where a respondent’s capacity or maximum performance capability is pitted against the extremity of difficulty of an item (Drasgow et al., 2010; Stark et al., 2006b). For example, in an ability test, a respondent with a high ability is expected to perform well because she is likely to dominate all the easy and moderately difficult items (i.e., getting them all correct), and get some of the hardest items correct.

However, when the studied field moves from ability to personality, the dominance models sometimes show inadequate fit. In fact, before Roberts et al. (2000) proposed the GGUM, several studies had already realized the unfolding property of some attitude statements (van Schuur & Kiers, 1994; Andrich, 1996; Roberts, Laughlin, & Wedell, 1999), which didn’t quite fit the single-peaked response function of dominance IRT models. One year after the GGUM was developed, Chernyshenko and colleagues (2001) fitted a variety of IRT dominance models (2PLM, 3PLM, and SGRM) to data from Goldberg’s Big Five Factor Markers (Goldberg, 1992), and the 16PF (Conn & Rieke, 1994). Surprisingly, all of the dominance IRT models showed misfit, and the chi-square fit statistics obtained were generally larger than those of cognitive ability tests. This is probably because that in personality tests, a different response process is applied which requires introspection (Chernyshenko et al., 2007). When people are considering personality items, they ask themselves “Does this statement closely describe me?” Therefore, the maximum probability of endorsement is achieved only when the item trait level matches the individual’s trait level, and the probability of endorsement decreases as the distance increases between the item and individual’s trait levels (Drasgow et al., 2010). This is the “unfolding technique” described by Coombs (1964), who coined the phrase “ideal point”. The unfolding property of items were proved by applying Levine’s (1984) maximum likelihood formula score model (MFSM). MFSM is a nonparametric IRT model, so it does not require an item to be logistic or monotonic to fit. It turned out that for item doubles and triples, MFSM showed better fit than the two logistic models, and more importantly, some of the items were found to have violated monotonicity, the hallmark of dominance models (Levine, 1984; Drasgow et al., 2010). Broadfoot (2008) showed that the GGUM had comparable fit with a partial credited model for conscientiousness and agreeableness data. Stark et al. (2006b) compared fitting of two ideal point models (GGUM and MSFM) with that of two dominance models (2PLM and MSFM with a dominance constraints) to data of 16PF (Conn & Reike, 1994). They concluded that ideal point models could fit personality items as well or better than dominance models, because they were able to fit both monotonic and non-monotonic items, the latter of which dominance models didn’t seem to handle well.

But the conclusion that in personality test, the ideal point model has better fit than the dominance model is not consistent across studies. Kosinski (2009) applied polytomous GGUM and SGRM to the Extraversion scale from the Goldberg’s 100-item Big Five personality questionnaire (Goldberg, Johnson, Eber, Hogan, Ashton et al., 2006), and found that GGUM had worse model-data fit than SGRM. Attempts to improve the fit by removing poor fit items were successful for SGRM but not for GGUM.

Researchers also got different results on the effects on model-data fit of non-monotonic, or intermediate items. For example, GGUM had worse fit than SGRM when there were no intermediate items on the test, and didn’t show significantly better model-data fit than SGRM until 50% of all items on the test were carefully selected, good intermediate items (i.e., items that have high α and close-to-zero δ under GGUM and low a-parameters under SGRM; Cao, Drasgow, & Cho, 2015). In a more recent study, Speer, Robie, and Christiansen (2016) fitted SGRM and GGUM to both both monotonic and non-monotonic conscientiousness and extraversion scales. They found that GGUM and SGRM fitted almost equally well for item singles, but that SGRM surpassed GGUM for items doubles and triples for all types of scales, even the non-monotonic scales.

Considering the inconsistent results and ongoing debate on the fit between the two types of IRT models and personality data, in the current study, we adopted both the dominance and the ideal point models. To be more specific, we chose the 2PLM and SGRM for the dominance IRT model, and polytomous and dichotomous GGUM for the ideal point model.

***DIF detection in the IRT framework***

In the current study, we utilized two paradigms to study DIF, and they are: (a) the null hypothesis significance testing (NHST) paradigm, and (b) the DIF effect size paradigm.

The NHST paradigm is the most popular approach to studying DIF. Under this paradigm, a null model and an alternative model are constructed and compared, and if the test statistic computed is statistically significant, then the studied item is considered a DIF item (Wang et al., 2013). We used two approaches to build the models: (a) the constrained baseline model, and (b) the free baseline model. We also chose the log-likelihood ratio (LR) as the test statistic for model comparison, because the LR test was shown to have yielded the best results in general (Wang et al., 2013).

***The Constrained Baseline Approach.*** The null model is constructed by constraining the paramters of all items to be equal across groups. A series alternative models are then constructed by freeing one item at a time. All alternative models are compared with the null model by comparing the log-likelihood, and the item truly has DIF when the alternative model has the greater log-likelihood and the difference of log-likelihood chi-square statistics exceeds a critical value (Wang et al., 2013). Due to the inflated Type I error rate of the constrained baseline approach (Stark et al., 2006a), if an item is considered free of DIF, then it’s safe to say that the item is a truly DIF-free. Such an item should be used as a linking item in the free baseline approach, so that across groups, the items will be put on the same scale. This is necessary given the fact that the measures in the current study are relatively short ones containing 20 items each (Lopez Rivas, Stark, & Chernyshenko, 2009).

***The Free Baseline Approach.*** The free baseline approach is preferred for detecting DIF items, because of its close-to-nominal Type I error rate (0.05) and high power with sample sizes as small as 250 (Lopez Rivas et al., 2009). Contrast to the constrained baseline approach, the free baseline model has a null model where the paramters of all items across groups are allowed to be freely estimated, except for those of the linking items. This model is constructed under the assumption that all non-linking items have DIF. Then a series of alternative models are constructed where non-linking items are constrained one at time, based on the assumption that the studied item has no DIF (Wang et al., 2013). The log-likelihood chi-square statistics are also obtained for model comparison, and an item has DIF if the log-likelihood of the null model is significantly greater than that of the alternative.

***The Log-likelihood Ratio Test.*** The LR test has been shown to be a good testing method for model comparison. In previous studies, the LR test was found to have high power for DIF detection (Wang, 2004; Stark et al., 2006a) and yield better results in general under GGUM, compared with other test methods such as the Akaike information criterion [AIC], Lord’s chi-square (Wang et al., 2013), and DFIT (Carter & Zickar, 2011). Therefore, in the current study, we adopted the LR test method for DIF detection.

***DIF Effect Size.*** Although long has been the most widely used and accepted test for hypothesis, NHST is believed to be limited and flawed. For example, NHST is thought to be trivial because the null hypothesis can always be shown to be false to some extent (Cohen, 1990), because an effect an always be found if measured accurately enough (Nye, 2011). Also, NHST is criticized for using a cutoff value to turn a continuum into a dichotomous reject/do not reject decision (Kirk, 2006; Nye, 2011). Another major limitation of NHST is that it provides little information on the magnitude, value, or importance of an effect (Kirk, 2006). It is possible that statistically significant DIF actually only has negligible effect size, especially when the sample size is large. LaPalme and colleagues (2016) found that 13 out of the 16 items contained significant DIF in the NHST paradigm, while according to DIF effect size, as many as 10 out of the 16 items had DIF that was too small to be meaningful (i.e., <.02; Cohen, 1992).

In order to obtain more accurate information on measurement non-equivalence, we included DIF effect size approach based on Nye (2011) in the current study. Nye’s DIF effect size method first computes the squared difference of between conditional expected scores (Wang et al., 2013), and then divides it by the pooled standard deviation of item *i* in the two groups (Nye, 2011). This second step puts the area difference on the standardized metric comparable to other effect size measures like Cohen’s *d*, and thus the DIF effect size can be interpreted the same way as how Cohen’s *d* is interpreted (Nye, 2011). Nye’s (2011) DIF effect size is given by:

,

where *fF(θ)* is the ability density of the focal group with the mean and variance estimated from the transformed **theta-hat** distribution (Nye, 2011).

**The Current Study**

The current study was designed to assess measurement equivalence of some facets of the Big Five Personality tests on the CPS with data collected from the U.S. and the mainland China.

Model-data fit was computed for the SGRM and the polytomous GGUM, and the source of misfit were explored by analyzing ICCs given by the two polytomous IRT models, as well the 2PLM and the dichotomous GGUM. The authors assessed DIF via the SGRM, the polytomous GGUM, and DIF effect size, in order to better understand the existence and effect of ME on the CPS acorss the U.S. and Chinese cultures.

**Method**

**2.1. Samples**

Data were collected from the United States and mainland China. 1183 American respondents finished the English-language version of the survey. 733 of them were undergraduate students from a large Midwestern university in the U.S., who enrolled in the study for course credit, and the rest were recruited from Amazon Mechanical Turk (MTurk). A total of 1654 Chinese undergraduate students from two universities in Nanjing, China took the Chinese-language of the survey.

3 quality control items were randomly embedded in the survey, and those who didn’t answer them all correctly got dropped from the analysis. We ended up with an American sample of 861 respondents (response rate = 72.78%; 66.5% females; mean age = 22.20 years; *SD* = 6.52). The racial makeup of the U.S. sample was 78.4% white, 7.8% African American, 6.4% Latino or Hispanic, 3.7% Asian, and 3.7% other. The final Chinese sample contained 1023 respondents (response rate = 61.85%; 82.7% females; mean age = 19.95 years; *SD* = 0.82).

**2.2. Measures**

In the current study, we assessed ME of the Well-being facet of Neuroticism, and the Curiosity facet of Openness from the CPS (Wang, 2013). We adopted a 4-point Likert-type scale, ranging from 1 (Strongly Disagree) to 4 (Strongly Agree), without a neutral response option. An undergraduate student from China studying at the University of Illinois translated the scale into Chinese. Both scales showed acceptable reliability in both groups (well-being: α = .852 for U.S., and α = .839 for Chinese; curiosity: α = .748 for U.S, and α = .783 for Chinese).

**2.3. Analysis**

Under the item response theory, both the dominance model and the ideal point model assumes unidimensionality, and therefore, we conducted an exploratory factor analysis (EFA) in SPSS to figure out data dimensionality. According to Reckase (1979), a scale is considered unidimensional if the first factor extracted accounted for at least 20% of the total variance. Results of principal axis factoring showed that both the well-being and the curiosity scales met the unidimensionality assumption. The percentages of total variance explained by the first factor extracted in the U.S./Chinses samples were 31.2%/29.1% for well-being, and 25.7%/34% for openness.

We first obtained GGUM item parameter estimates in GGUM2004 (**Roberts et al., 2000**) for both groups and both scales, respectively, because the GGUM does not require reverse coding. Item parameter estimates and responses were then thrown in the Modfit software (**Stark, 2007**) to examine model-data fit based on the sample-size adjusted chi-square to degrees of freedom ratio obtained from the software for item singles, doubles, and triples. MODFIT generated the item characteristic curves (ICCs) at the same time, which were used to determine which items were supposed to be reversely coded before any analyses could be conducted under the dominance model. The SGR model item parameters were estimated in MULTILOG 7.0 software (Thissen, Chen, & Bock, 2003). Model-data fit for the SGR model was also computed using MODFIT. Adequate fit is indicated by Chi-square-to-degree-of-freedom ratios less than 3 (Tay, Ali, Drasgow, & Williams, 2011). Source of misfit was also explored by assessing the ICCs of potential intermediate items, both under polytomous and dichotomous IRT models.

DIF NHST was conducted using a combination of the constrained and free baseline model approach. The constrained baseline model approach was first used to find DIF-free items, which were used as linking items in the free baseline model. The constrained baseline model is more conservative in detecting DIF-free items due to the inflated Type I error rate (Stark et al., 2006a), while the free baseline model is more effective in finding DIF items, because of the low Type I error rate and high power (Lopez Rivas et al., 2009). The log-likelihood ratio statistic was used for NHST, based on a previous finding (Wang et al., 2013) that the LR test performs consistently well with different types of data. DIF effect size was also computed based on Nye (2011) as an implement to the NHST to provide more specific information on DIF magnitude.