**2. 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. 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 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 facets met the unidimensionality assumption. The percentages of total variance explained by the first factor extracted in the U.S./Chinses samples are 31.2%/29.1% (well-being), and 25.7%/34% (openness).

We first obtained GGUM item parameter estimates in GGUM2004 (**Roberts et al., 2000**) for both groups and both scales, respectively. Item parameter estimates and responses were then thrown in MODFIT (**Stark, 2007**) to examine model-data fit based on the sample-size adjusted chi-square to degrees of freedom ratio obtained from the software. MODFIT generated the item characteristic curves (ICCs) at the same time, which were used to determine which items were negative and thus should have been reversely coded before any analyses were conducted under the dominance model. The model used for the dominance model was Samejima’s Graded Response (SGR) model, and item parameters were estimated in MULTILOG (**citation, 1000**). Model-data fit for the SGR model was also computed using MODFIT. Model fit for item singles, doubles, and triples were all computed and investigated. Adequate fit is indicated by Chi-square-to-degree-of-freedom ratios less than 3 (Tay et al., 2011). Source of misfit was also explore 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 (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 implement to the NHST to provide more specific information on DIF magnitude and effect.

1. Introduction

It is widely believed that personality traits are important, for they have been shown to correlate with a variety of important constructs, including XXXXXXXXX (a lot of citations on criterion-related validity of personality traits). Therefore, it is important to also examine measurement equivalence of personality measures in a cross-cultural setting.

***Overall:***

1. Cross-cultural personality tests – ME – non-IRT methods (FA): factor or scale level; we need item-level DIF analysis – IRT methods
2. IRT methods: model choice important – polytomous; dominance model (SGR) or the ideal point model (GGUM)
3. SGR is predominant; but why GGUM – Drasgow et al., 2000 (including introduction of GGUM and SGR; assumptions, equations etc.)
4. DIF: Wei and others simulation results comparing SGR and GGUM 🡪 GGUM more advantageous
5. However, mixed results:

**Wei 2016 paper:**

intended to compare SGR and GGUM using WLEIS EI scale, but GGUM fitted so badly that they had to drop the method – they believed that it was because WLEIS purported to measure EI as an ability rather than a personality trait

but other EI studies have showed that GGUM can fit adequately to EI scale (Zampetakis, 2011)

**Speer et al. (2016) paper, consistent with Kosinski (1999):**

Used extraversion and conscientiousness scales; GGUM fitted better for item singles only with non-monotonic scales; showed severe misfit for item doubles and triples, with both the non-monotonic and monotonic scales.

**Kosinski (1999) paper:**

GGUM showed worse model-data fit than SGR for both the original and optimized 20-item IPIP extraversion scale. GGUM model-data fit also failed to improve after the optimization, while SGR showed great model-data fit improvement. He believed that the misfit of GGUM may have been due to the fact that the extraversion scale contained highly extreme items. This potential source of misfit can be detected by looking at the fit plots, despite the fact that Chernyshenko et al. (2007) found that GGUM fits monotonic items equally well as SGR does.

**Cao et al. (2015)**

Consistent with Kosinski (1999), GGUM was also found to have fitted worse than SGRM when there’s no intermediate items on the scale (Cao et.al., 2015). GGUM didn’t show significantly better model-data fit than SGRM until 50% of all items were carefully selected working intermediate items (i.e., items that have high α and close-to-zero δ under GGUM and low a-parameters under SGRM).

1. There are two types of intermediate items, one that looks like one, and one that also works like one.
2. No study so far has (successfully) analyzed empirically DIF under GGUM and SGRM on personality tests.
3. DIF effect size: also necessary, because NHST is unable to show DIF magnitude. Some significant DIF may not be large enough to be meaningful. (Cite Nye paper)
4. CPS: introduction (Wei, 2013).