MEASUREMENT EQUIVALENCE OF THE COMPREHENSIVE PERSONALITY SCALES ACROSS CULTURES: AN ITEM RESPONSE THEORY APPROACH

BY

LUYAO ZHANG

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, 2017

Urbana, Illinois

Adviser:

Professor Fritz Drasgow

ABSTRACT

Within the item response theory (IRT) framework, this study compared cross-culturally different approaches to the assessment of differential item functioning (DIF) in two personality tests of the Comprehensive Personality Scale (Wang, 2013). A dominance IRT model (SGRM) and an ideal point model (the GGUM) were applied within the NHST paradigm, due to the debate over which is the more appropriate model for personality research. Nye's (2011) DIF effect size measure was also used in the current study to overcome the oversensitivity of NHST to large sample size. Participants from the U.S. (n = 861) and China (n = 1023) responded to two personality scales from the CPS: the Well-being scale, and the Curiosity scale. Results indicated that SGR was applicable for DIF assessment, but the NHST paradigm was so sensitive to large samples that even trivial DIF could be significant. GGUM failed to work in the DIF analyses due to ill-conditioned matrices. The DIF effect size measure compensated for the NHST method by providing the magnitude of DIF. Implications for future research and practice are discussed.

献给爸爸妈妈。

To my parents.

ACKNOWLEDGEMENTS

First, I would like to thank my adviser, Professor Fritz Drasgow, for his feedback on this thesis, and his continuous support of my research. He allowed this paper to be my own work, but steered me in the right direction whenever I needed it.

I would also like to thank the second author of this paper, Dr. Liwen Liu, for keeping me sane throughout this journey.

Last but not least, I would like to acknowledge Professor Hua-hua Chang as the second reader of this thesis, and I am very grateful for his valuable comments.

This couldn't have happened without you all.

Table of Contents

CHAPTER 1

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, 2002), 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, their 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 an observed mean score difference is due to true group differences or to relationships that vary across groups between the latent variable and the observed scores (Raju, Laffitte, & Byrne, 2002). According to Drasgow (1984), ME is obtained when participants from different groups have the same expected observed score as long as they were at the same latent trait level. Testing for ME in cross-cultural personality tests is essential given measurement non-equivalence has been found in items on a variety of cross-cultural personality tests, including 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 that we always assess ME before scores are compared across groups or any selection decisions are made based upon these scores.

The two major approaches to the study of ME are Confirmatory Factor Analytic (CFA) mean and covariance structure (MACS) analysis, 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 levels 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 tested. Configural invariance is the weakest type of ME, and it tests for the existence of the same number of factors and similar loading patterns across groups. Metric invariance refers to factor loadings being invariant across groups. Scalar invariance, the strongest form of invariance of the three, implies that, in addition to metric invariance, when items are regressed on latent variables, they have the same intercepts across groups (Steenkamp & Baumgartner, 1998).

The alternative approach to studying ME is IRT-based differential item functioning (DIF). DIF is different from the CFA approach in several ways. First, CFA tests the three different types of ME one after the other, 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, differential test functioning (DTF) in the IRT context takes into consideration the possible compensatory nature of DIF (Raju, van der Linden, & Fleer, 1995; Raju et al., 2002), an issue that's rarely 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 can help us to identify the source of DIF (LaPalme, Wang, Joseph, Saklofske, & Yan, 2016). Lastly, within the IRT framework, we can assess DIF using an ideal point model, which some previous studies have found to be more appropriate for self-report 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 for IRT analysis. It derives from Likert's (1932) approach to analyzing rating scales, and assumes that the higher a participant's trait level, the more likely she will answer positively. But it doesn't mean that the ideal point model should 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 responses 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 DIF studies for personality tests 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 proceeded with only the dominance model because of the severe misfit of the Generalized Graded Unfolding Model (GGUM), a type of ideal point model that has been widely used. 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 for personality tests under both the dominance and the ideal point model, but the comparison was between a group of applicants and a group of incumbents, rather than groups from different cultures. In Carter, Dalal, Zickar, and Adams (2009), the DIF approach was applied under the GGUM to examine the effects of vague quantifiers, 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 used a DIF effect size measure (Nye, 2011) in our study.

In summary, the current study intended to examine measurement equivalence of some scales of the Comprehensive Personality Scale (CPS; Wang, 2013) via an IRT DIF method. The analysis was done across 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 with both models, and DIF effect sizes were computed to obtain DIF magnitude. Finally, we evaluated the existence and effects of intermediate items on model-data fit through 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.

Different Assumptions Underlying the Dominance and the Ideal Point Models

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 for analyzing 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 recently. 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:

$$P_i(\theta) = \frac{\exp[Da_i(\theta - b_i)]}{1+\exp[Da_i(\theta - b_i)]},$$

where $P_i(\theta)$ is the probability of a random respondent correctly answering Item i correctly.

There are two item parameters in a 2PLM. a_i is the discrimination parameter that represents the degree to which an item separates latent adjacent trait levels (Maurer, Raju, & Collins, 1998). The larger a_i is, the steeper the IRF will be. b_i 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).

Samejima's (1969) Graded Response (SGR) model is an extension 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 functions (BRF), which are obtained by successively merging response options (Kosinski, 1999). The probability of a respondent with a trait level equal to θ selecting response option *k* equals the probability of endorsing response option *k* and higher minus that of endorsing response option *k*+1 and higher. The probability of selecting option *k* on item *i* is given by:

$$P_{i,k}(\theta) = \frac{1}{1+\exp\left[-Da_i\left(\theta - b_{i,k}\right)\right]} - \frac{1}{1+\exp\left[-Da_i\left(\theta - b_{i,k+1}\right)\right]}$$

The item paramters (a_i, $b_{i,k}$) and scaling constant (D) mean 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. Among the few ideal point models, the most employed is the the Generalized Graded Unfolding Model (GGUM; Roberts et al., 2000), which is applicable to both dichotomous and polytomous response data. As discussed above, ideal point models assume a response process different from dominance models. The 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:

$$P\left[Z_i = z \mid \theta_j\right] = \frac{\exp\{\alpha_i[z(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\} + \exp\{\alpha_i[(M-z)(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}]\}}{\sum_{w=0}^{C}\{\exp\{\alpha_i[w(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}]\} + \exp\{\alpha_i[(M-w)(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}]\}\}}$$

This function gives the probability associated with the *j*th respondent's observable response to the *i*th item. Z_i 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. α_i is the discrimination parameter, and δ_i is the location parameter of item i on the latent trait continuum. τ_{ik} 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 τ_{ik}s are symmetric about the point (θ_j - δ_i) = 0.

Model-Data 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). 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 (Drasgow et al., 2010).

However, when the studied field moves from ability to personality, the dominance models sometimes show inadequate fit. In fact, before the GGUM (Roberts et al., 2000) was developed, 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 monotonically increasing response function of dominance IRT models. One year after the GGUM was proposed, Chernyshenko and colleagues (2001) fitted a variety of IRT dominance models (2PLM, 3PLM, and SGRM) to data obtained using 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 seen for cognitive ability tests. This was probably because in personality tests, a different response process was applied which requires introspection (Chernyshenko et al., 2007). To be more specific, 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 was 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 the fit to data of the 16 PF (Conn & Reike, 1994) for two ideal point models (GGUM and MSFM) with that for two dominance models (2PLM and MSFM with a dominance constraints). They concluded that ideal point models could fit personality items as well or even 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 tests, 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 obtained different results on the model-data fit effects of non-monotonic, or intermediate items. For example, GGUM had worse fit than SGRM when there were no intermediate items on the test, and did not 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 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 over 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 to represent the dominance IRT models, and polytomous and dichotomous GGUM to represent the ideal point models.

DIF detection in the IRT framework

In the current study, we utilized two paradigms to study DIF: (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 approach, and (b) the free baseline approach.

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 parameters of all items to be equal across groups. A series of 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 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 other items can 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). In 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 model.

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, 2011b). 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 paradigm for testing hypotheses, NHST is 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), and an effect can always be found if the sample is large 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), with both samples of more than 500 people, 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 the DIF effect size approach based on Nye (2011) in the current study. Nye's DIF effect size method first computes the mean squared difference 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), thus putting the area difference on the standardized metric comparable to other effect size measures like Cohen's *d*. The pooled standard deviation is given by:

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

Therefore, the DIF effect size can be interpreted the same way as Cohen's *d* is interpreted (Nye, 2011). The DIF effect size is given by:

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

where $f_F(\theta)$ is the ability density of the focal group with the mean and variance estimated from the transformed $\hat{\theta}$ distribution (Nye, 2011).

The Current Study

The current study was designed to assess measurement equivalence of some facets of 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 was 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 across the U.S. and Chinese cultures.

CHAPTER 2

METHOD

Samples

Data were collected from the United States and the 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.

Three quality control items were randomly embedded in the survey, and those who didn't answer them all correctly were 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).

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). Responses were made using a 4-point Likert-type scale, ranging from 1 (Strongly Disagree) to 4 (Strongly Agree), without a neutral response option. Two undergraduate student from China studying in the United States translated the scales into Chinese, and two back translated. Both scales showed acceptable reliability in both groups (Well-being: $\alpha = .852$ for the U.S. group, and $\alpha = .839$ for the Chinese group; Curiosity: $\alpha = .748$ for the U.S group, and $\alpha = .783$ for the Chinese group).

Analyses

Both the dominance model and the ideal point model assume unidimensionality, and therefore, we conducted an exploratory factor analysis (EFA) in SPSS to examine 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 Curiosity.

We first obtained GGUM item parameter estimates with the GGUM2004 software (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 analyzed with the MODFIT software (Stark, 2007) to assess model-data fit based on the sample-size adjusted chi-square to degrees of freedom ratio computed for item singles, doubles, and triples. MODFIT generated the item characteristic curves (ICCs) at the same time, which were used to determine which items should be reverse coded before any analysis could be conducted with the dominance model. After negative items were reversed, the SGR model item parameters were then estimated with 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). Sources of misfit were 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 its 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 the 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 a complement to the NHST for information on DIF magnitude.

CHAPTER 3

RESULTS

Model Fit

We examined the model-data fit for the GGUM first, because unlike SGR, GGUM does not require reverse coding. Based on the ICCs given by MODFIT, we discarded items with flat characteristic curves in at least one of the groups, because they had poor discrimination and contained little information. Also based on the ICCs were decisions about which items to be reverse coded for the dominance models. If, as the latent trait level went up, the probability of the participants endorsing the item went down, then the item was considered a negative item, and reversed.

The Well-being scale. Based on the ICCs, Items 6, 19, and 20 were excluded from further analyses because of low discrimination. More specifically, Items 6 and 20 were not discriminating enough for the Chinese group, while Item 19 had flat ICCs in the U.S. group. Among the remaining 17 items, 9 were reversed for both groups based on the ICCs as well as the loadings given by a one-factor CFA. Model fit was then obtained using theses 17 items for both GGUM and SGR, with negative items reversed for the latter. Results of the model fit analyses can be found in Table 1.

Adequate fit is indicated by Chi-square-to-degree-of-freedom ratios less than 3 (Drasgow, Levine, Tsien, Williams, & Mead, 1995; Tay et al., 2011). Based on this criterion, as shown in Table 1, both GGUM and SGR exhibited good fit for item singles, but some misfit for item doubles and triples. In the current study, we focused on the fit of item doubles and triples. This is because that item singles are insensitive to misfit when item parameters and fit are computed using the same sample (Drasgow et al., 1995).

Item doubles and triples have been found to be sensitive to local dependence, and for a 17-item scale measuring a specific personality facet, local dependence is not rare (Chernyshenko et al., 2007), so a higher cutoff for misfit may be more proper (Speer et al., 2016). Also, if there's misfit for more than one model, relative misfit of the two models can still be compared (Stark et al., 2006b), and as shown in Table

1, for item singles, GGUM fitted slightly better than SGR in the U.S. group, while in the Chinese group, the two models showed equally good fit. In both groups, GGUM fitted better than SGR for item doubles. For item triples, in the U.S. group, SGR fitted only faintly better than GGUM, while in the Chinese group, GGUM fitted better than SGR.

Considering that in general, for the Well-being scale, polytomous GGUM had better model fit than SGR, and that both models showed acceptable, if not satisfactory fit, we decided to keep both models for the DIF analyses.

We believed that the source of the worse model fit for SGR was the unfolding items on the scale (Stark et al., 2006b). Unfolding items are non-monotonic, and thus violate the assumption of monotonicity underlying SGR and other dominance IRT models. GGUM, assuming non-monotonicity, is capable of modeling unfolding items and thus take advantage of the unfolding property of the item. To identify unfolding items, we went back to the ICCs and item parameter estimates, and noticed one item: Item 17 ("I am positive, but negative thoughts can conquer me sometimes"). Under GGUM, in both groups, Item 17 had discrimination parameters that were not large, yet acceptable (U.S: 0.82; CH: 0.83) and location parameters close to zero (U.S: -0.22; CH: -0.66). Moreover, across the two groups, a lot of the response option functions for this item were bell-curved (Figures 1-2). These characteristics are what one should expect from an item that is working as an unfolding/intermediate item. Another characteristic of an unfolding item is that it won't be modeled very well by the dominance model, because of the non-monotonicity. Sure enough, by examining the ICCs (Figures 3-4) and item parameters of Item 17 under SGR, we found that this model was unable to capture the unfolding property, producing minimal discrimination parameters (U.S.: 0.09; CH: 0.06), and extremely large difficulty parameters (U.S.: -20.67; CH: -43.52). To further assess the effects of Item 17 on model fit and relative model fit, we computed new model fit without Item 17 for the two models (Table 2). As expected, without the unfolding item, the model fit of SGR now became almost as good as GGUM, mainly due to the significant improvement of the model fit of SGR.

In order to examine the unfolding item more closely, we tried intensifying the unfolding pattern by having fewer response option functions (ROF) for each item (i.e., dichotomizing the response data). We went through the exact same process as with polytomous data, starting from examining model-data fit under GGUM with all 20 items. The only difference was that this time we kept Item 19, which was dropped before for low discrimination. Items 6 and 20 were deleted as under polytomous data, due to low discrimination. Model-data fit with these 18 items for both GGUM and 2PLM was computed, which can be found in Table 3. As shown in Table 3, both GGUM and 2PLM exhibited much better fit than with polytomous data. All combinations of group, model, and item types demonstrated adequate or almost adequate fit, except for item triples for the U.S. group under 2PLM, which showed some misfit, but nothing severe. Same as when with polytomous data, GGUM fitted generally better than 2PLM across two groups.

Item 17 was again identified via GGUM ICCs and item parameter estimates as the only unfolding item. Under GGUM, the unfolding property of Item 17 was demonstrated through the large discrimination parameters (U.S.: 1.88; CH: 1.41), close-to-zero location parameters (U.S.: -0.01; CH: -0.39), and steep bell-curved ICCs (Figures 5-6). 2PLM, similar to SGR, failed to model the unfolding item by having near zero discrimination (U.S.: 0.05; CH: 0.01), extremely large difficulty parameters (U.S.: -15.25, CH: -74.35) and flat ICCs' (Figures 7-8). When Item 17 was dropped, the model fit of 2PLM for item doubles and triples in both groups improved by more than 30% (Table 4), while the improvement for GGUM was trivial.

The Curiosity scale. Item 1 was dropped before any analyses were carried out due to translation error. Items 10 and 12 were also dropped, because no participants endorsed "Strongly disagree", which is a situation that GGUM2004 couldn't deal with without combining response options. But we were unable to combine the responses of these two items, because MODFIT couldn't handle scales with inconstant numbers of response categories. However, items having an option that no one endorsed was no problem for Multilog, so we kept these two items for analyses under SGR. We also excluded Items 9, 16, and 19 from further analyses due to low discrimination in at least one group. To be more specific, Items 9 and 16

had low discrimination parameters for the U.S. group, and all 3 items had flat ICCs in the Chinese group. Model fit was then computed under GGUM with the remaining 14 items, and under SGR with 17 items (Items 10 and 12 were kept). Table 5 contains the model-data fit results. Both models showed some misfit for item doubles and triples across groups. Compared with SGR, GGUM showed worse fit in the U.S. group, but slightly better fit in the Chinese group.

Given the fact that the data-model fit was not too bad, we decided to include both models in our DIF analyses.

By examining the GGUM item parameters and ICCs, in the Chinese group, we were able to identify Item 13 ("I am as curious as anybody else I know") as a weak non-monotonic item with a pretty low discrimination parameter (0.29), close-to-zero location parameter (-0.69), and bell-curved option response functions (Figure 9) for two of the response categories. The same item, under SGR, had option response functions (Figure 10) that were rather flat, a close-to-zero a-parameter (0.06), and an extreme b-parameter (-34.31). In the U.S group, however, no item showed identifiable non-monotonicity. All items had location parameters that were very far away from 0, demonstrating monotonicity rather than non-monotonicity. Item 13 had similar ICCs under GGUM and SGM in the U.S. group (Figures 11-12).

After Item 13 was removed, we recomputed model-data fit (Table 6). As shown in Table 6, GGUM still fitted worse than SGR for the U.S group, but for the Chinese group, SGR now fitted almost as well as GGUM, mainly because the model fit of GGUM got worse after the unfolding item was removed.

Next, we dichotomized the response data for a clearer view of the unfolding item. 19 items were used (Item 1 was dropped due to inaccurate translation). Items 9, 13, and 16 showed poor discrimination, and thus were deleted. Item 13 was a weakly non-monotonic item under polytomous GGUM for the Chinese group. Interestingly, this time, Item 19 exhibited non-monotonicity. Note that Item 19 was deleted under polytomous GGUM due to low discrimination for the Chinese group. Under polytomous GGUM, although Item 19 had poor discrimination for the Chinese group, it was in fact non-monotonic in the U.S. group (Figure 13).

Therefore, model-data fit was computed in MODFIT without Items 1, 9, 13, and 16 (see Table 7). Again, dichotomous IRT models had much better fit than their polytomous counterparts, with all fit indices smaller than 3, indicating excellent fit. The GGUM fitted a little better than 2PLM. Item 19 was identified as an unfolding item in both groups under GGUM (Figures 14-15), with acceptable yet not large discrimination parameters (U.S.: 0.63; CH: 0.58), and close-to-zero location parameters (U.S.: 0.17; CH: -0.07). ICCs (Figures 16-17) of the same item under 2PLM showed that the model did not capture the non-monotonicity as well as dichotomous GGUM, but the general misfit was not worth worrying about. This was probably because that Item 19 was not that discriminating even under GGUM. When Item 19 was removed, fit became almost equally good for both models.

DIF

The Well-being scale. With the constrained baseline approach under polytomous GGUM, when we freed a different item each time, GGUM2004 reported multiple times that the matrices were too ill-conditioned and thus the inverse may have been inaccurate. Being unable to obtain trustworthy linking items, we turned to ICCs and effect sizes, and were able to identify at least one item as the linking item for the free baseline analysis. However, during the free baseline analysis, under polytomous GGUM, many of the matrices again turned out to have been too ill-conditioned to produce accurate results. Therefore, we had to drop the GGUM from our DIF analyses.

Table 9 presents the DIF results obtained with SGR, and Nye's DIF effect size measure for the Well-being scale. Items 6, 19, and 20 were dropped from the analysis because they had low discrimination. Under SGR, all items had significant DIF according to the constrained baseline approach, and thus the item with the smallest negative twice the difference between log-likelihood after and before it was freed (31.8; critical value with Bonferroni correction: 16.06; $df = 4$) was chosen as the linking item for the free baseline approach. The free baseline approach, with an ideal Type I error rate, also identified all the non-linking items as DIF items. Therefore, all items were flagged as DIF items under SGR using a NHST paradigm. However, based on Cohen's (1992) guidelines for interpreting effect size, 4 out of the 17 items showed a negligible DIF effect size smaller than .2 (Items 3, 9, 16, and 17), 2 items exhibited moderate

DIF (i.e., $.5 \leq d < .8$; Items 5 and 7), only 2 items exhibited large DIF (i.e., $0.8 \leq d$; Items 10 and 15), and the remaining 9 items showed small DIF ($.2 \leq d < .5$).

Although we were not able to compute DIF for the polytomous GGUM, we examined differential test functioning (DTF) by combining the test characteristic curves (TCC) of the two groups. As shown in Figure 18, under GGUM, the scale exhibited very small DTF, as the two pretty straight TCCs almost completely overlapped. DTF under SGR (Figure 19), on the other hand, was larger and non-uniform. To be more specific, when the well-being level was below 0, the Chinese participants had lower expected total score, whereas these scores became higher than the U.S. participants when the trait level was above 0. The two TCCs were very slightly S-shaped under SGR.

The Curiosity scale. For the Curiosity scale, with both baseline approaches, many of the matrices were reported to have been ill-conditioned, so eventually we had to exclude the GGUM from the analyses again.

As shown in Table 10, when SGR was applied, except for Item 12, all the other items were found to have significant DIF. With the exact same set of items, according to DIF effect sizes, 4 out of the 17 items had negligible DIF (Items 2, 5, 7, and 20), 4 exhibited small DIF (Items 3, 6, 8, and 12), only 2 showed large DIF (Items 11 and 17), and the other 6 items exhibited moderate DIF.

TCCs for the Curiosity scale were also computed for the two groups, and were combined to examine DTF (Figures 20-21). In general, under GGUM, the scale showed smaller DTF than under SGR, and DTF under both models was non-uniform. The two TCCs crossed at almost the same trait level (i.e., approximately -2.25) under the two models. Below this trait level, the Chinese participants had very slightly lower expected total scores than the American participants, but the trend reversed past this point, and the differences in the expected scores became larger.

CHAPTER 4

DISCUSSION

The current study revealed that GGUM and SGR had comparable model fit with the CPS data. For the Well-being scale, with as few as just one unfolding item showing properties of a good non-monotonic item, GGUM managed to achieve equally good or better fit than SGR across the two groups. For the Curiosity scale, GGUM fitted better than SGR in the Chinese group, while worse than SGR in the U.S. group. Removal of the unfolding item greatly improved the model fit of SGR. Although the model fit of GGUM improved slightly when the unfolding item was removed, the extent was a lot less than that of SGR. These findings were inconsistent with Speer et al. (2016), where SGR surpassed GGUM for items doubles and triples for all types of scales, even the non-monotonic scales. In their study, whether an item was non-monotonic was based on expert ratings. Since no ICCs or item parameters were presented in the Speer et al. (2016), we doubt whether items rated high in non-monotonicity actually worked as unfolding items (i.e., having unfolding ICCs, acceptable alpha paramters, and close-to-zero delta parameters). According to the current study, having more working unfolding items will likely to harm the fit for SGR, but not or not as much as for GGUM.

The current study also demonstrated that the SGR, as a dominance IRT model, is applicable for DIF analysis of personality tests in a cross-cultural setting. However, when the sample size is large as in our study, the NHST paradigm became so sensitive that even a small DIF could lead to rejection to the null hypothesis. As a result, given by the log-likelihood ratio test, all items on the Well-being scale and 15 out of 16 items on the Curiosity scale were identified to have significant DIF across the two groups. When DIF effect sizes were examined with Nye's (2011) method, not surprisingly, the Well-being and Curiosity scales each had only 2 items with large DIF, while all the other items flagged DIF under NHST only had small to moderate DIF.

The fact that in our study, both scales were shown to have smaller DTF under GGUM than SGR, points out the importance of considering the use of GGUM in DIF analyses using personality data.

Apparently in the current study, according to GGUM, both the Well-being and the Curiosity scales are more equivalent cross cultures than when they are examined with SGR.

Due to ill-conditioned matrices, we were not able to carry out the DIF analysis with constrained or free baseline approach under GGUM. This was because results obtained from ill-conditioned matrices could not be trusted. We were unable to find any studies mentioning getting ill-conditioned matrix warnings with GGUM2004, but problems have been reported of having singular or invertible matrices with the model, which led to the authors deleting those data sets from their simulation study (Carter & Zickar, 2011a). We would like to point out that, in GGUM2004, the warning for an ill-conditioned matrix will only appear in a command window, and stay on for about 2 seconds before the window closes normally. No warnings will be shown in the GGUM2004 output file, and all results, including the fit indices and time spent carrying out the analysis will be computed as usual. Therefore, when an automated program (e.g., constrained baseline models) is left running unsupervised in GGUM2004, it is possible that the results are inaccurate because of ill-conditioned matrices, yet the executive output still looks normal. We suggest that researchers supervise the whole process, and examine the item parameter standard errors carefully. Standard errors that are too large or too close to zero may indicate ill-conditioned matrices, but not necessarily. GGUM2004 should be improved upon so that output files could include warnings about singular or ill-conditioned matrices. Having to drop GGUM from the current study was disappointing.

Future research

In the future, simulation studies should be carried out, with the hope of identifying the factors that may cause singular or ill-conditioned matrices, and exploring solutions for such conditions other than simply giving up the model. This will be particularly important for studies using real data, where it's almost impossible to delete the problematic data sets and proceed with the normal ones.

More attention should be paid to applying the GGUM to real data, especially personality data obtained cross-culturally, rather than only to simulation studies. Moreover, how and why various types of unfolding, or intermediate items work or not work should be closely examined, given the importance of such items to the fit of the GGUM, which was demonstrated in the current study.

Conclusion

Although by applying the dominance IRT model and NHST, we found significant DIF on almost all items on the Well-being and Curiosity scales of the CPS, DIF effect size measures told a different story by demonstrating that only 2 items on each scale had large DIF. Also, contrasted to LaPalme et al. (2016), and Speer et al. (2016), we found that the GGUM fitted better or almost as well as the dominance model, which is in line with previous studies advocating the application of GGUM in personality research (e.g., Drasgow et al., 2010; Stark et al., 2006b).

TABLES AND FIGURES

Table 1
Model fit for polytomous GGUM and SGR for the Well-being scale (Item 6, 19, and 20 dropped).

Adjusted χ^2 /*df* ratios

Models, sample, and items	Frequency of χ^2 /*df*				Mean	SD
	<1	1<2	2<3	>=3		
GGUM						
US						
Singles	17	0	0	0	0.001	0.003
Doubles	5	14	20	97	6.383	5.189
Triples	3	22	69	586	5.908	3.497
CHN						
Singles	17	0	0	0	0	0
Doubles	10	9	14	103	6.241	5.541
Triples	2	24	64	590	6.062	3.342
SGR						
US						
Singles	17	0	0	0	0.024	0.1
Doubles	7	9	23	97	7.016	6.763
Triples	3	27	77	573	5.891	3.792
CHN						
Singles	17	0	0	0	0	0
Doubles	6	9	18	103	6.555	6.757
Triples	0	17	66	597	6.429	3.824

Note: Good model-data fit is indicated by χ^2/df smaller than 3. Both groups showed good fit for item singles, but some misfit for item doubles and triples. However, considering the prevalence of local dependence, the misfit is not too severe.

Table 2
Model fit for polytomous GGUM and SGR for the Well-being scale (Items 6, 17, 19, and 20 dropped).

Adjusted χ^2 /*df* ratios

Models, sample, and items	Frequency of χ^2 /*df*				Mean	SD
	<1	1<2	2<3	>=3		
GGUM						
US						
Singles	16	0	0	0	0.001	0
Doubles	7	11	17	85	5.998	5.052
Triples	4	24	60	472	5.595	3.434
CHN						
Singles	16	0	0	0	0	0
Doubles	9	10	12	89	5.885	4.419
Triples	2	14	49	495	5.827	2.768
SGR						
US						
Singles	16	0	0	0	0.016	0.062
Doubles	7	10	21	82	5.628	4.631
Triples	3	29	77	451	5.142	3.136
CHN						
Singles	16	0	0	0	0	0
Doubles	6	10	17	87	5.546	4.168
Triples	1	18	62	479	5.539	2.607

Note: Good model-data fit is indicated by χ^2/df smaller than 3. Both groups showed good fit for item singles, but some misfit for item doubles and triples. However, considering the prevalence of local dependence, the misfit is not too severe.

Table 3
Model fit for dichotomous GGUM and 2PLM for the Well-being scale (Items 6 and 20 dropped).

Adjusted χ^2 /*df* ratios						
Models, sample, and items	Frequency of χ^2 /*df*				Mean	SD
	<1	1<2	2<3	>=3		
GGUM						
US						
Singles	18	0	0	0	0	0
Doubles	118	9	3	23	1.383	3.89
Triples	452	92	66	206	2.315	3.982
CHN						
Singles	18	0	0	0	0	0
Doubles	116	11	7	19	1.482	4.274
Triples	445	114	65	192	2.273	3.7
2PLM						
US						
Singles	18	0	0	0	0	0
Doubles	114	6	5	28	2.298	8.717
Triples	379	103	65	269	3.956	8.301
CHN						
Singles	18	0	0	0	0	0
Doubles	110	11	9	23	1.968	6.784
Triples	390	108	73	245	3.185	5.837

Note: Good model-data fit is indicated by χ^2/df smaller than 3. Under GGUM, both groups showed good fit, while under 2PLM, there was slight misfit for item triples.

Table 4
Model fit for dichotomous GGUM and 2PLM for the Well-being scale (Items 6, 17, and 20 dropped).

Adjusted χ^2 /*df* ratios						
Models, sample, and items	Frequency of χ^2 /*df*				Mean	SD
	<1	1<2	2<3	>=3		
GGUM						
US						
Singles	17	0	0	0	0	0
Doubles	100	9	7	20	1.361	3.329
Triples	349	95	60	176	2.281	3.742
CHN						
Singles	17	0	0	0	0	0
Doubles	106	8	6	16	1.336	4.079
Triples	378	106	60	136	2.101	3.625
2PLM						
US						
Singles	17	0	0	0	0	0
Doubles	104	6	6	20	1.429	4.045
Triples	366	87	55	172	2.369	4.069
CHN						
Singles	17	0	0	0	0	0
Doubles	105	8	6	17	1.359	4.105
Triples	377	98	63	142	2.122	3.629

Note: Good model-data fit is indicated by χ^2/df smaller than 3. Under both models, the two groups showed adequate fit.

Table 5
Model fit for polytomous GGUM and SGR for the Curiosity scale (Items 1, 9, 10, 12, 16, and 19 dropped for GGUM; Items 1, 9, 16, and 19 dropped for SGR).

Adjusted χ^2 /*df* ratios

Models, sample, and items	Frequency of χ^2 /*df*				Mean	SD
	<1	1<2	2<3	>=3		
GGUM						
US						
Singles	13	1	0	0	0.081	0.303
Doubles	4	3	5	79	8.373	6.039
Triples	2	5	16	341	8.073	4.045
CHN						
Singles	14	0	0	0	0	0
Doubles	3	9	7	72	5.876	3.959
Triples	2	8	31	323	5.684	2.533
SGR						
US						
Singles	16	0	0	0	0.05	0.172
Doubles	11	8	16	85	6.484	5.07
Triples	3	27	51	479	6.188	3.129
CHN						
Singles	16	0	0	0	0.029	0.115
Doubles	7	6	7	100	6.528	3.864
Triples	1	4	31	524	6.303	2.445

Note: Good model-data fit is indicated by χ^2/df smaller than 3. Both groups showed good fit for item singles, but some misfit for item doubles and triples under the two models. In general, the Chinese group had less severe misfit than the U.S. group.

Table 6
Model fit for polytomous GGUM and SGR for the Curiosity scale (Item 1, 9, 10, 12, 13, 16, and 19 dropped for GGUM; Items 1, 9, 13, 16, and 19 dropped for SGR).

Adjusted χ^2 /*df* ratios						
Models, sample, and items	Frequency of χ^2 /*df*				Mean	SD
	<1	1<2	2<3	>=3		
GGUM						
US						
Singles	13	0	0	0	0.05	0.18
Doubles	4	3	5	66	7.901	6.204
Triples	2	4	19	261	7.478	3.981
CHN						
Singles	13	0	0	0	0	0
Doubles	2	6	4	66	6.216	4.01
Triples	1	1	12	272	6.216	2.479
SGR						
US						
Singles	15	0	0	0	0.005	0.019
Doubles	10	10	16	69	5.948	5.003
Triples	4	26	54	371	5.57	2.904
CHN						
Singles	15	0	0	0	0.026	0.1
Doubles	6	6	4	89	6.553	3.911
Triples	0	4	16	435	6.58	2.505

Note: Good model-data fit is indicated by χ^2/df smaller than 3. Both groups showed good fit for item singles, but some misfit for item doubles and triples. However, considering the prevalence of local dependence, the misfit is not too severe.

Table 7
Model fit for dichotomous GGUM and 2PLM for the Curiosity scale (items 1, 9, 13, and 16 dropped for both models).

Adjusted χ^2 /*df* ratios						
Models, sample, and items	Frequency of χ^2 /*df*				Mean	SD
	<1	1<2	2<3	>=3		
GGUM						
US						
Singles	16	0	0	0	0	0
Doubles	96	6	6	12	0.912	2.748
Triples	340	84	49	87	1.486	2.49
CHN						
Singles	16	0	0	0	0	0
Doubles	80	8	6	26	1.628	2.832
Triples	211	88	72	189	2.658	2.835
2PLM						
US						
Singles	16	0	0	0	0	0
Doubles	98	3	7	12	0.916	2.851
Triples	341	82	48	89	1.506	2.527
CHN						
Singles	16	0	0	0	0	0
Doubles	81	9	4	26	1.646	2.913
Triples	204	83	77	196	2.741	2.874

Note: Good model-data fit is indicated by χ^2/df smaller than 3. In both groups, adequate model fit is demonstrated across item singles, doubles, and triples.

Table 8
Model fit for dichotomous GGUM and 2PLM for the Curiosity scale (items 1, 9, 13, 16, and 19 dropped).

Adjusted χ^2 /*df* ratios						
Models, sample, and items	Frequency of χ^2 /*df*				Mean	SD
	<1	1<2	2<3	>=3		
GGUM						
US						
Singles	15	0	0	0	0	0
Doubles	83	5	4	13	1.021	2.941
Triples	260	70	43	82	1.656	2.617
CHN						
Singles	15	0	0	0	0	0
Doubles	70	7	6	22	1.696	2.997
Triples	164	63	65	163	2.822	2.978
2PLM						
US						
Singles	15	0	0	0	0	0
Doubles	84	3	6	12	1.016	2.993
Triples	261	71	41	82	1.652	2.634
CHN						
Singles	15	0	0	0	0	0
Doubles	70	8	3	24	1.709	3
Triples	162	66	64	163	2.854	2.981

Note: Good model-data fit is indicated by χ^2/df smaller than 3. In both groups, adequate model fit is demonstrated across item singles, doubles, and triples.

Table 9
DIF results obtained via SGR and DIF effect size for the Well-being scale (Items 6, 19, and 20 were deleted).

		SGR	DIF Effect Size
Item 1	I often feel depressed.	O	0.467
Item 2	I sometimes find myself thinking negative thoughts.	O	0.324
Item 3	I would say that I am happy more often than most other people.	O	0.172**
Item 4	I tend to react negatively to life events.	O	0.304
Item 5	I'm unhappy that life is unfair to me.	O	0.569
Item 7	I feel I am always treated unfairly.	O	0.686
Item 8	I easily feel discouraged.	O	0.285
Item 9	I am about as well off in life as most people.	O	0.095**
Item 10	Usually I will not allow negative thoughts to occupy my mind for a long time.	O	0.823
Item 11	I believe I can lead a life without regrets.	O	0.381
Item 12	I feel confident about my ability to do most things.	O	0.499
Item 13	It is difficult for me to make through misfortunes in my life.	O	0.424
Item 14	Bad things happen in life, but I can handle it pretty well.	O	0.361
Item 15	I see difficulties all around me.	O	1.029
Item 16	I tend to be a pessimistic person.	O	0.175**
Item 17	I am positive, but negative thoughts can conquer me sometimes.	O	0.028**
Item 18	I would consider myself an optimistic person.	O	0.262

Note: O = the item has significant DIF. ** = the DIF effect size is smaller than 0.2, and thus should be considered not different from a DIF-free item.

Table 10
DIF results obtained via SGR and DIF effect size for the Curiosity scale.

		SGR	DIF Effect Size
Item 2	I learn new things only when I have to.	O	0.112**
Item 3	I am not really interested in new technology.	O	0.307
Item 4	I am usually intrigued by what I learn in classes.	O	0.643
Item 5	I only care about information that is relevant to me.	O	0.079**
Item 6	I sometimes try new things just so I can learn more about them.	O	0.270
Item 7	I can be persuaded to try some new things, but most of the time I am reluctant to do so.	O	0.142**
Item 8	I sometimes read non-fiction books to learn something new.	O	0.322
Item 10	I am interested in what is happening around the world.	O	0.616
Item 11	I am excited about new knowledge.	O	1.108
Item 12	I like to learn new things whenever I have time.	X	0.323
Item 13	I am as curious as anybody else I know.	O	0.603
Item 14	I am not curious about the things that I don't know.	O	0.717
Item 15	I would prefer a job where I don't have to learn anything new.	O	0.656
Item 17	I am fascinated by science.	O	1.246
Item 18	I am not interested in learning new things.	O	0.635
Item 20	I try new restaurants only when other people recommend them.	O	0.114**

Note: O = the item has significant DIF. X: the item does not have significant DIF. ** = the DIF effect size is smaller than 0.2, and thus should be considered not different from a DIF-free item.

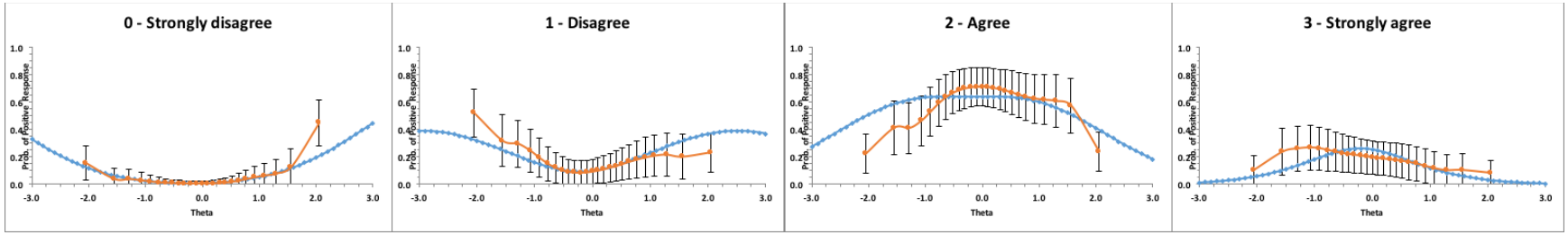

Fig. 1. IRT item characteristic curves of Item 17 under GGUM for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for the standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, GGUM was able to capture, although not perfectly, the unfolding property of the data.

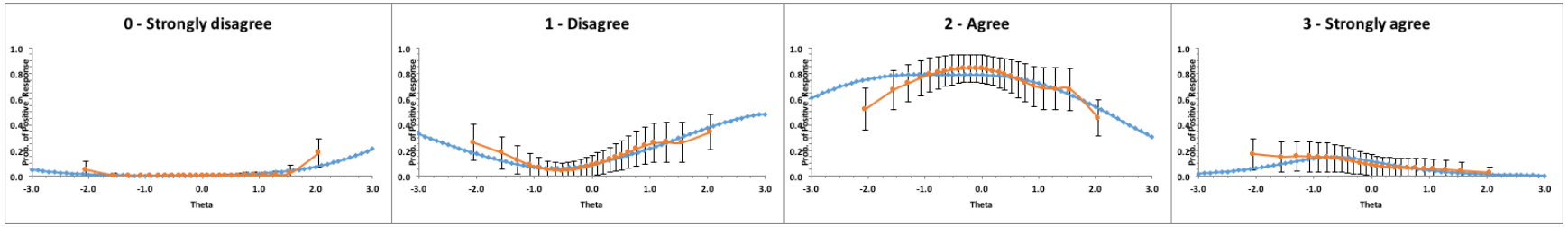

Fig. 2. IRT item characteristic curves of Item 17 under GGUM for the Chinese group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for the standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, GGUM was able to model the unfolding pattern of the data.

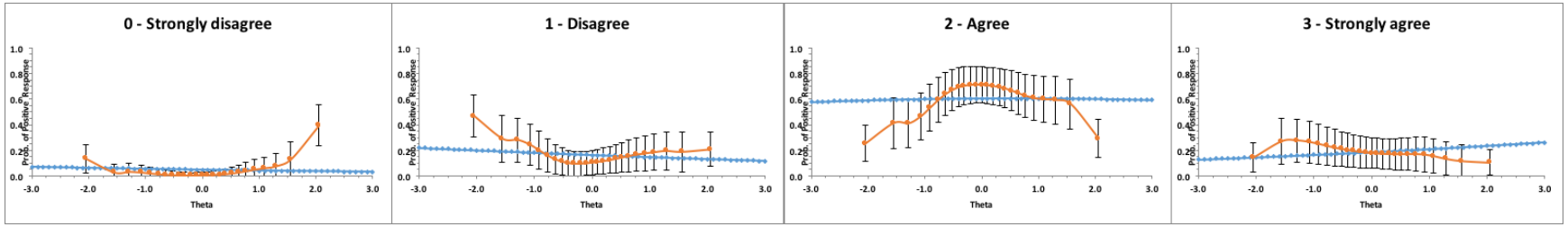

Fig. 3. IRT item characteristic curves of Item 17 under SGR for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, SGR failed to capture the unfolding characteristic of the data.

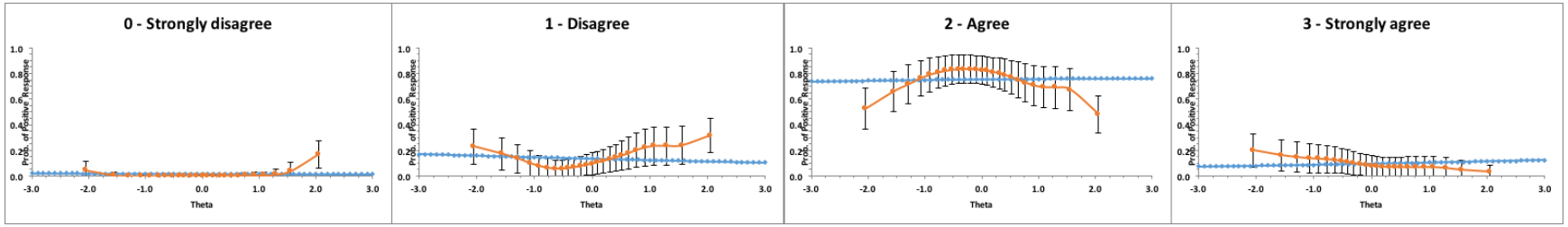

Fig. 4. IRT item characteristic curves of Item 17 under SGR for the Chinese group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, SGR failed to capture the unfolding characteristic of the data, especially for "Disagree", and "Agree". As shown in the plots, SGR failed to capture the unfolding characteristic of the data.

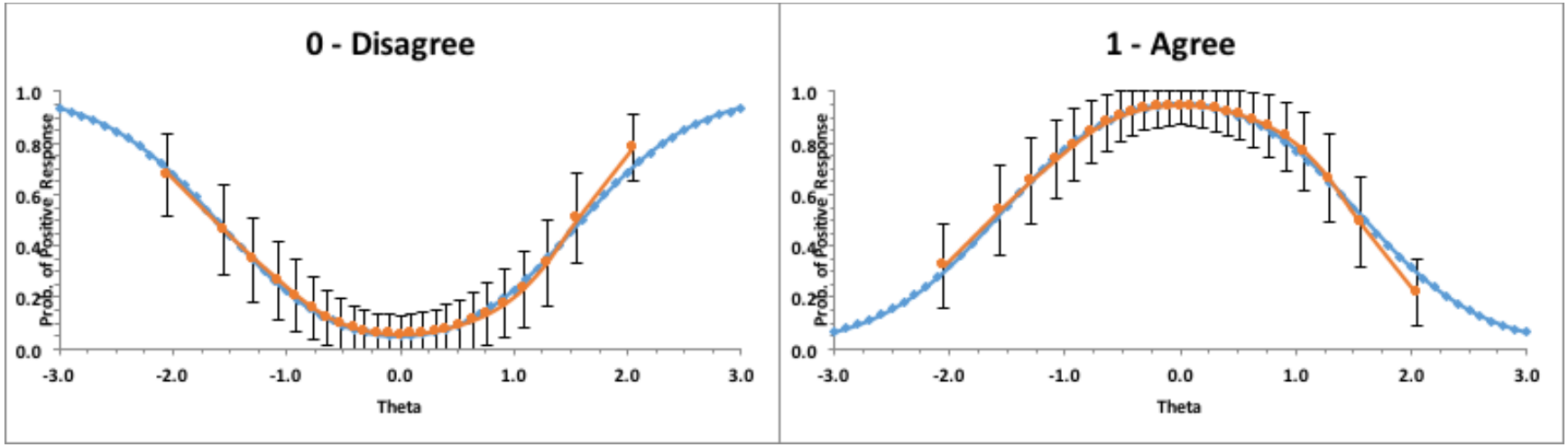

Fig. 5. IRT item characteristic curves of Item 17 under dichotomous GGUM for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. In the plots, the responses consist of 0 (disagree), and 1 (agree). As shown in the plot, the GGUM successfully captured the unfolding pattern shown by the data (i.e., the orange lines). As shown in the plots, GGUM modelled almost perfectly the unfolding characteristic.

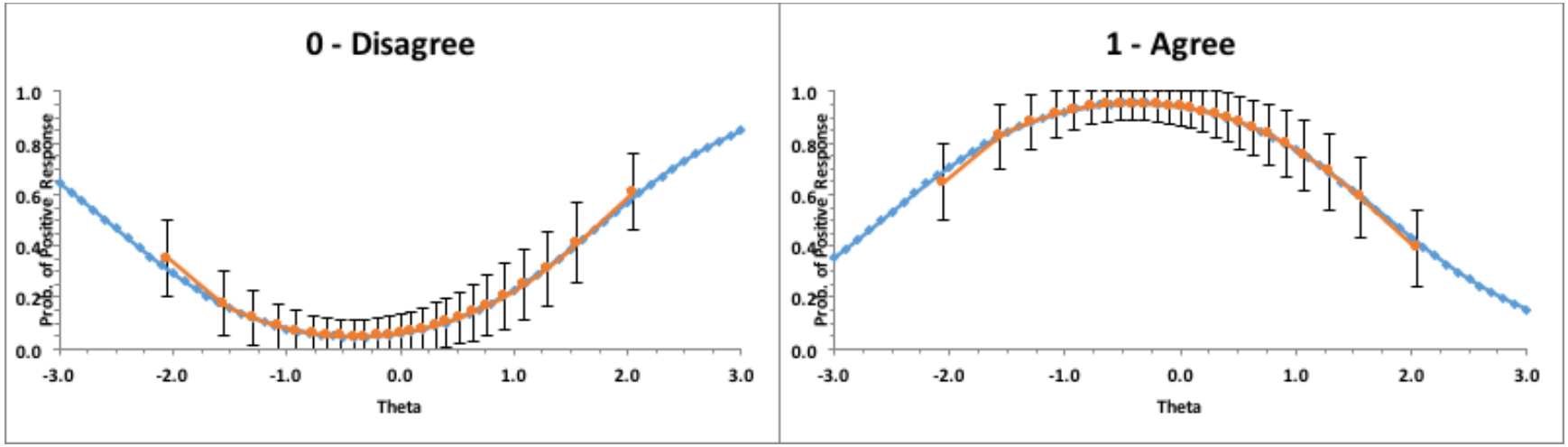

Fig. 6. IRT item characteristic curves of Item 17 under dichotomous GGUM for the Chinese group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. In the plots, the responses consist of 0 (disagree), and 1 (agree). As shown in the plots, the unfolding characteristic was captured accurately by GGUM.

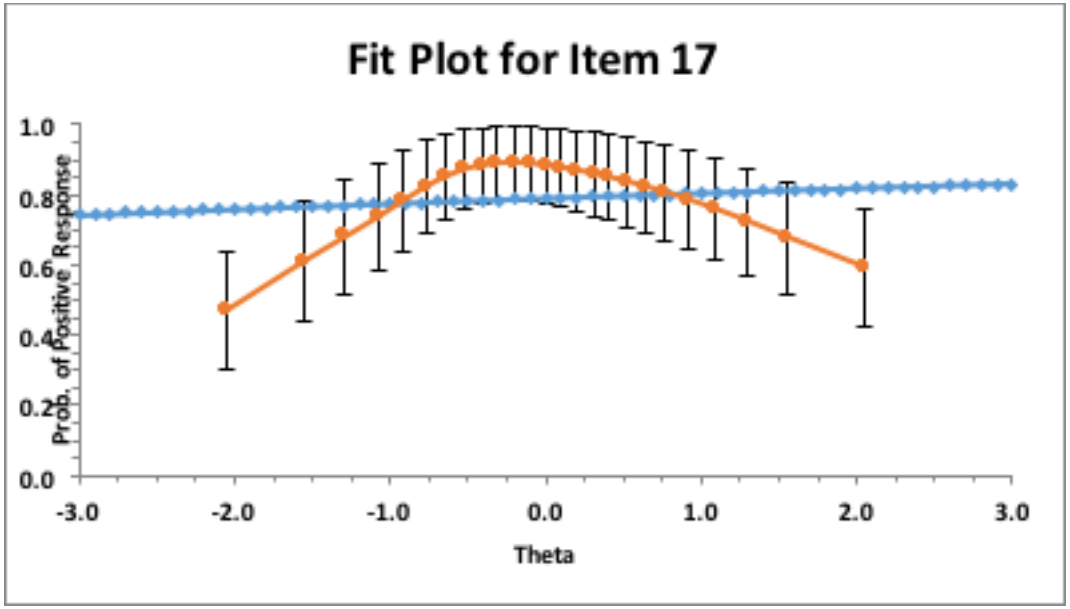

Fig. 7. IRT item characteristic curves of Item 17 under 2PLM for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. In the plots, the responses consist of 0 (disagree), and 1 (agree). As shown in the plot, the dominance model was unable to model the unfolding pattern given by the data (i.e. the orange line). As shown in the plot, the straight response function failed to model the unfolding pattern.

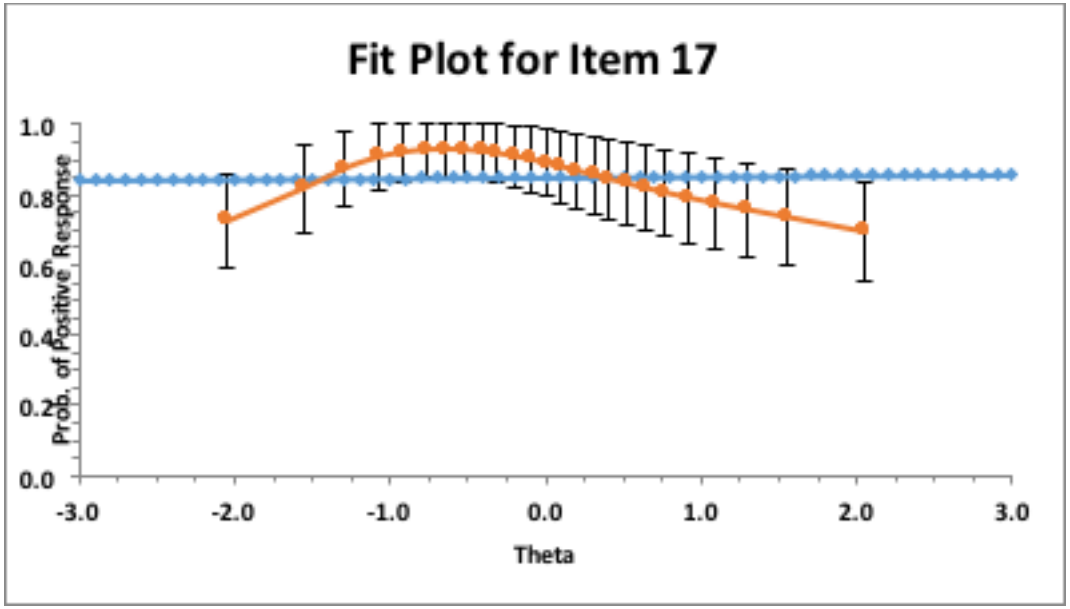

Fig. 8. IRT item characteristic curves of Item 17 under 2PLM for the Chinese group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. In the plots, the responses consist of 0 (disagree), and 1 (agree). As shown in the plot, the dominance model was unable to model the unfolding pattern given by the data (i.e. the orange line). As shown in the plot, the misfit came from the dominance model's failure to model the unfolding characteristic.

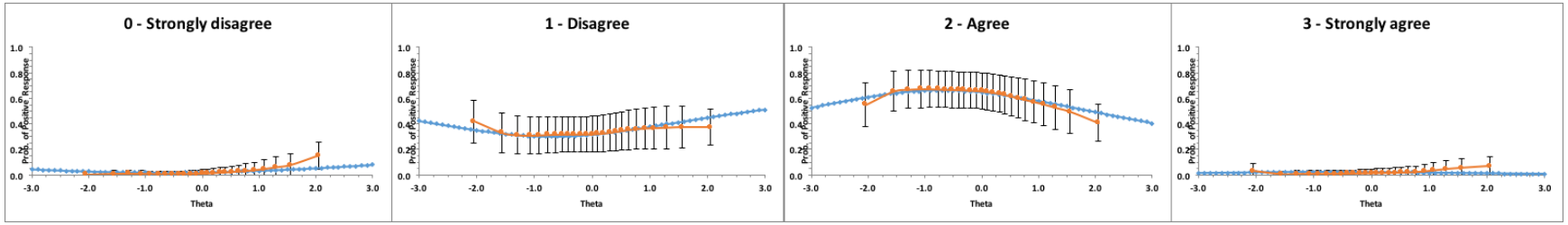

Fig. 9. IRT item characteristic curves of Item 13 under polytomous GGUM for the Chinese group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, SGR failed to capture the unfolding characteristic of the data, especially for "Disagree", and "Agree". As shown in the plots, GGUM was able to model the unfolding pattern.

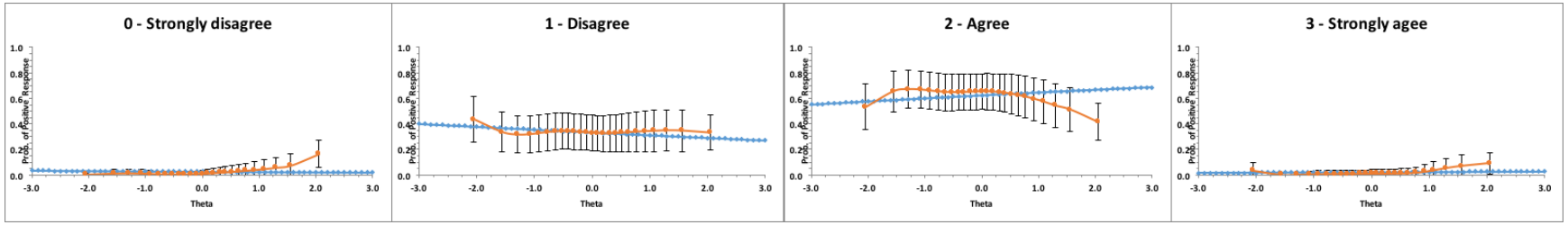

Fig. 10. IRT item characteristic curves of Item 13 under SGR for the Chinese group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, SGR failed to capture the unfolding characteristic of the data, especially for "Disagree", and "Agree". As shown in the plots, SGR did not do a good job modelling the unfolding item.

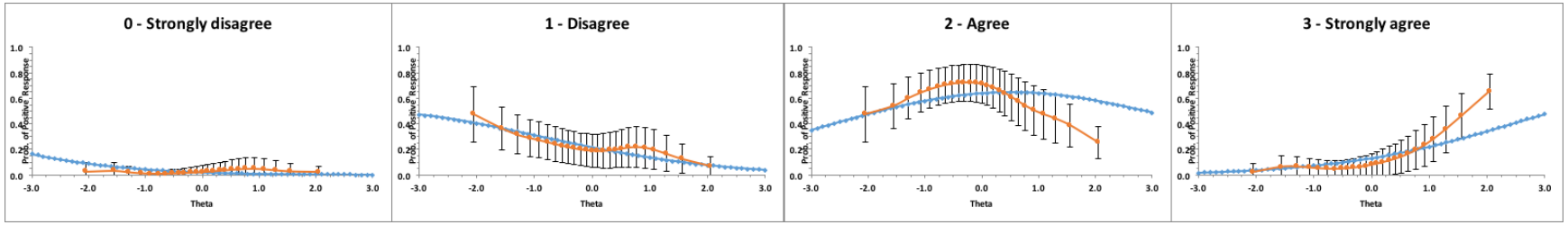

Fig. 11. IRT item characteristic curves of Item 13 under polytomous GGUM for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, SGR failed to capture the unfolding characteristic of the data, especially for "Disagree", and "Agree".

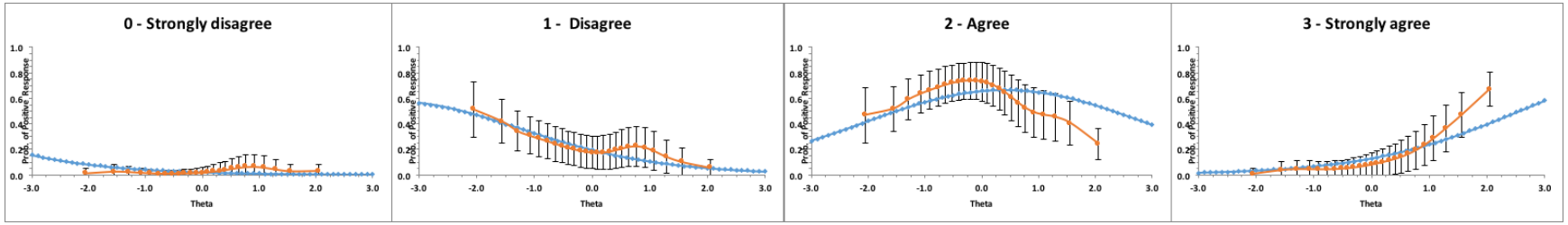

Fig. 12. IRT item characteristic curves of Item 13 under SGR for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, SGR failed to capture the unfolding characteristic of the data, especially for "Disagree", and "Agree".

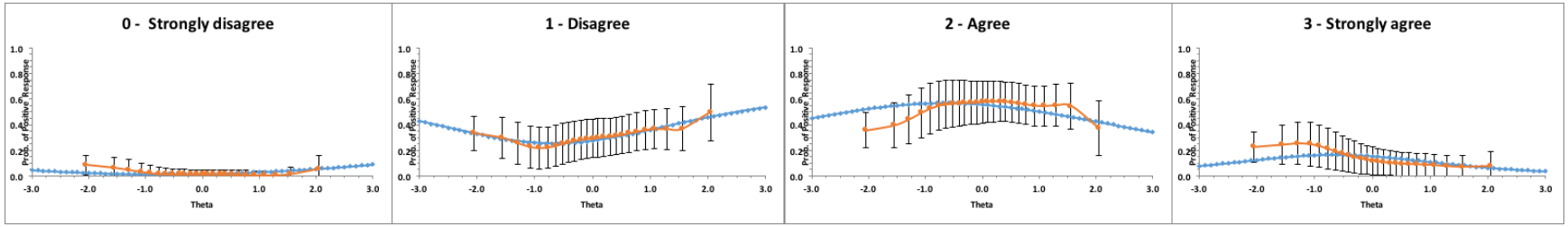

Fig. 13. IRT item characteristic curves of Item 19 under polytomous GGUM for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. Respondents responded to the survey on a scale consisting of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), but MODFIT requires that responses start from 0, and thus in the plots, the responses consist of 0 (strongly disagree), 1 (disagree), 2 (agree), and 4 (strongly agree). As shown in the plots, SGR failed to capture the unfolding characteristic of the data, especially for "Disagree", and "Agree". According to the plots, the item is an unfolding item, and GGUM was able to capture the characteristic.

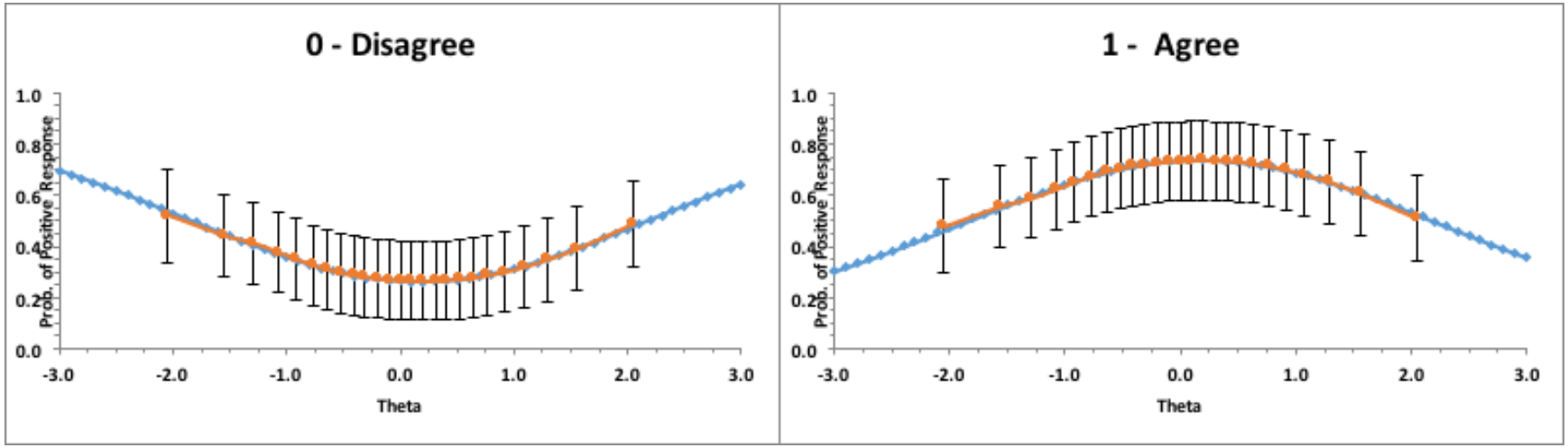

Fig. 14. IRT item characteristic curves of Item 19 under dichotomous GGUM for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. In the plots, the responses consist of 0 (disagree), and 1 (agree). As shown in the plots, the unfolding characteristic was captured accurately by GGUM.

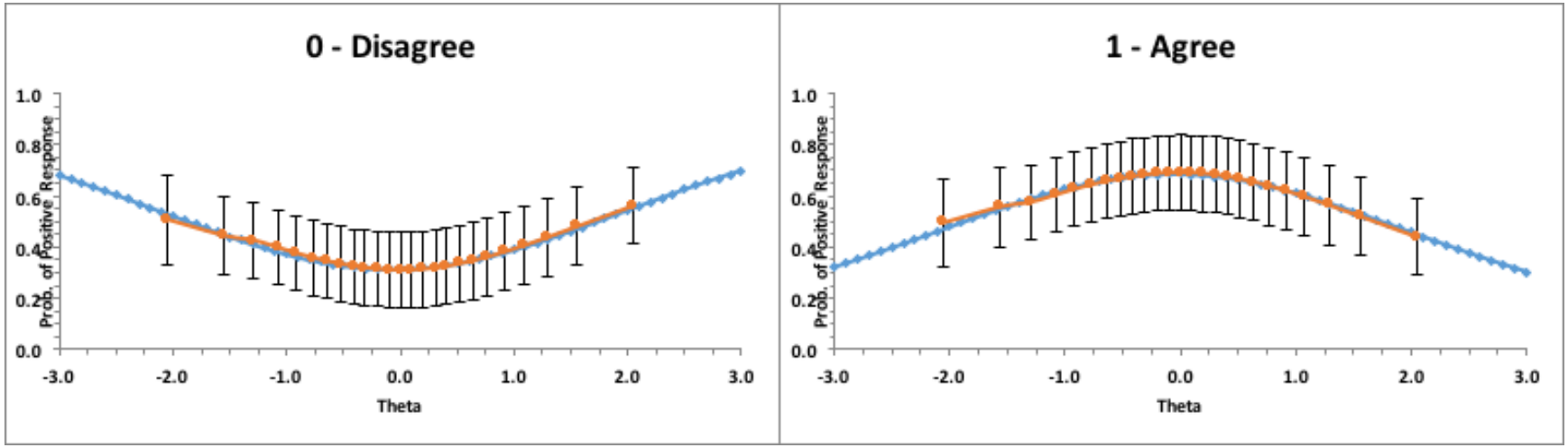

Fig. 15. IRT item characteristic curves of Item 19 under dichotomous GGUM for the Chinese group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. In the plots, the responses consist of 0 (disagree), and 1 (agree). As shown in the plots, the unfolding characteristic was captured well by GGUM.

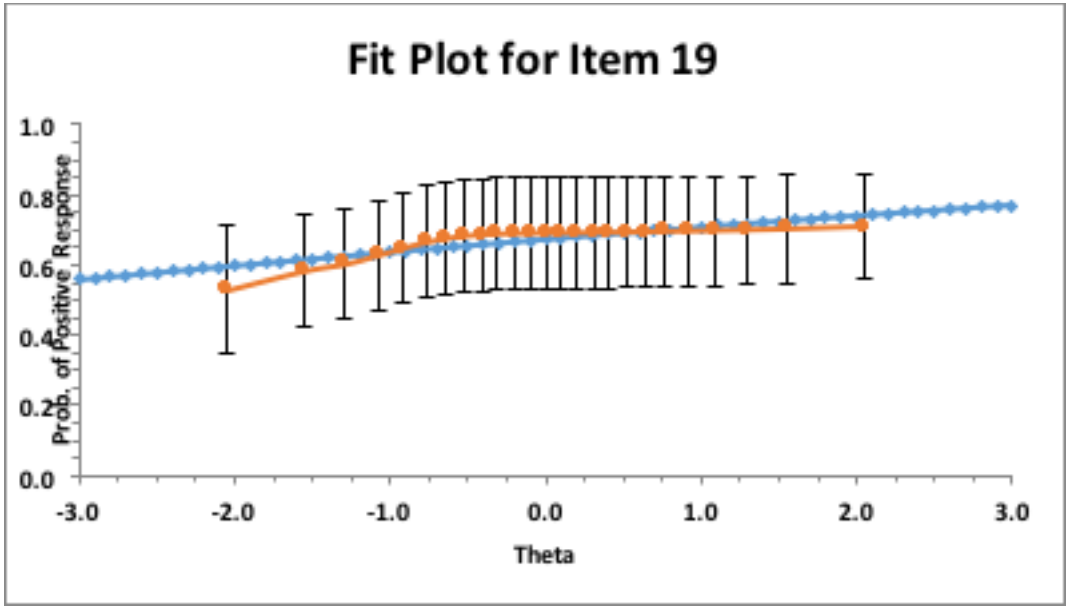

Fig. 16. IRT item characteristic curves of Item 19 under 2PLM for the U.S. group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. As shown in the plot, 2PLM failed to model the unfolding pattern.

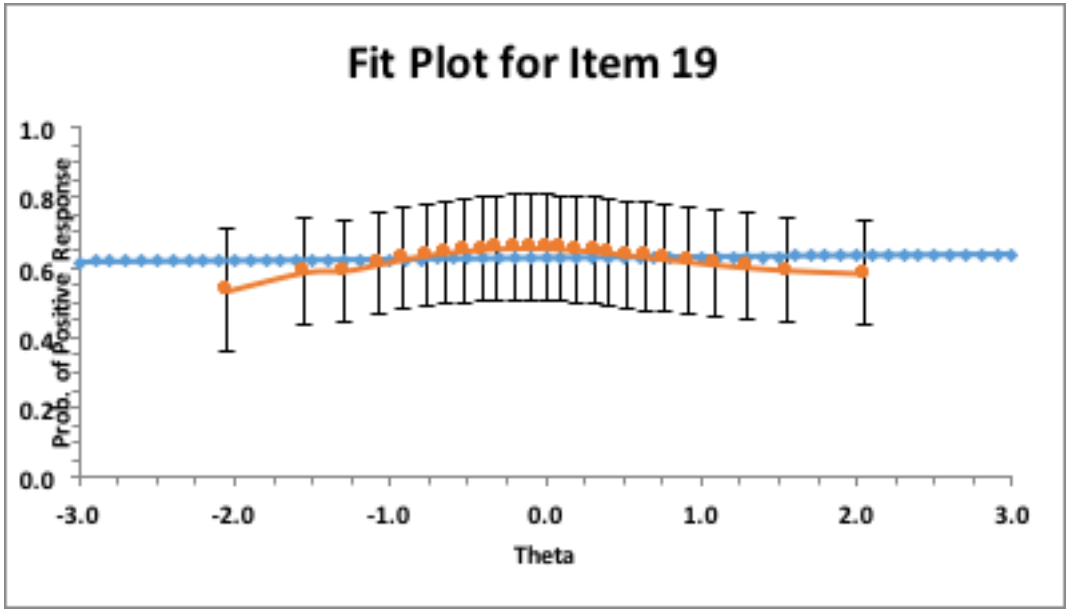

Fig. 17. IRT item characteristic curves of Item 19 under 2PLM for the Chinese group. Note: the orange line represents the empirical response function, and the blue line represents the response function that derives from the model. Vertical bars stand for standard error. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the probability of positive responses from 0 to 1. In the plots, the responses consist of 0 (disagree), and 1 (agree). As shown in the plot, 2PLM failed to model the weak unfolding characteristic.

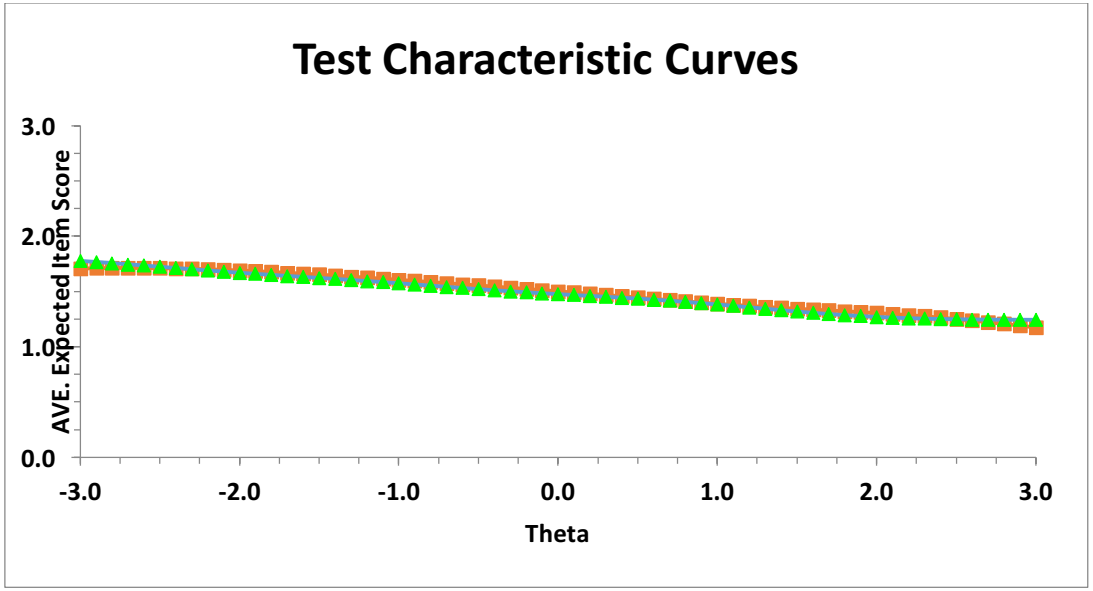

Fig.18. IRT test characteristic curves (TCCs) of the Well-being scale under polytomous GGUM for the U.S. and the Chinese groups. Note: the red line represents the TCC of the Chinese group, and the green line represent the TCC of the U.S. group. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the expected item score.

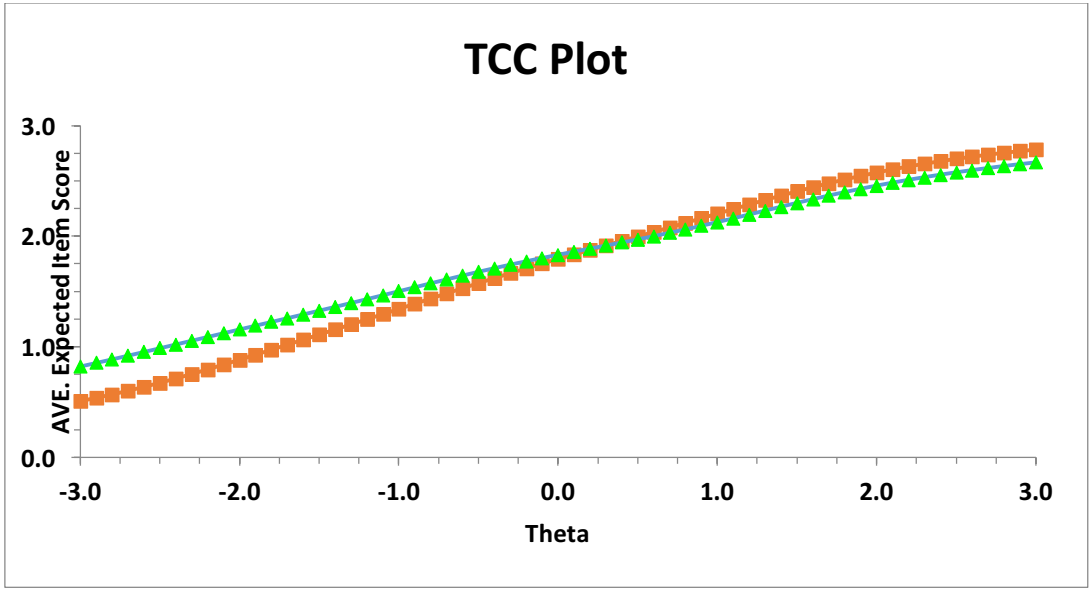

Fig.19. IRT test characteristic curves (TCCs) of the Well-being scale under SGR for the U.S. and the Chinese groups. Note: the red line represents the TCC of the Chinese group, and the green line represent the TCC of the U.S. group. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the expected item score.

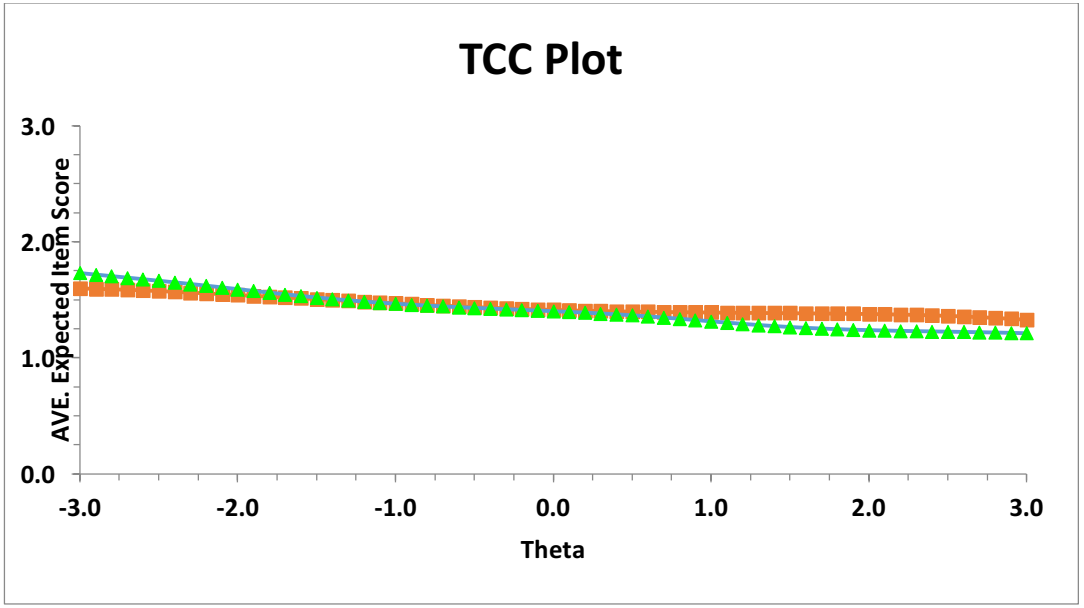

Fig.20. IRT test characteristic curves (TCCs) of the Well-being scale under polytomous GGUM for the U.S. and the Chinese groups. Note: the red line represents the TCC of the Chinese group, and the green line represent the TCC of the U.S. group. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the expected item score.

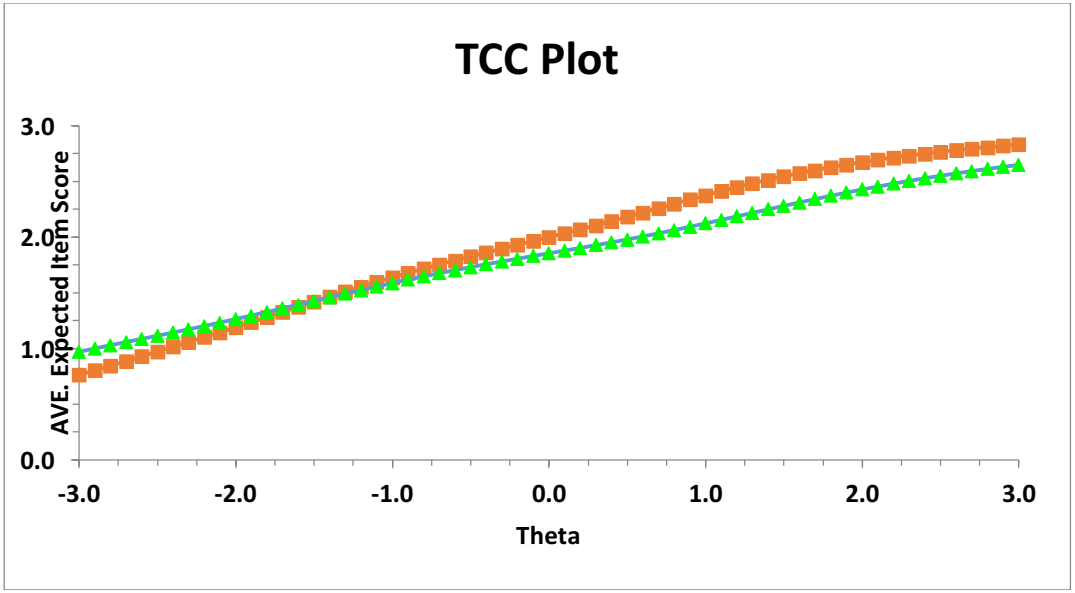

Fig.21. IRT test characteristic curves (TCCs) of the Well-being scale under SGR for the U.S. and the Chinese groups. Note: the red line represents the TCC of the Chinese group, and the green line represent the TCC of the U.S. group. The horizontal axis "Theta" represents the latent continuum from -3.0 to +3.0, and the vertical axis represents the expected item score.

REFERENCES

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12*(1), 33-51. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/617442870?accountid=14553

Baranik, L. E., Lakey, C. E., Lance, C. E., Hua, W., Meade, A. W., Hu, C., & Michalos, A. (2008). Examining the differential item functioning of the Rosenberg self-esteem scale across eight countries. *Journal of Applied Social Psychology,38*(7), 1867-1904. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/622038181?accountid=14553

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/617946789?accountid=14553

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. International Journal of Selection and Assessment, 9(1-2), 52-69. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1111/1468-2389.00163

Broadfoot, A. A. (2008). *Comparing the dominance approach to the ideal-point approach in the measurement and predictability of personality* Available from PsycINFO. (621760437; 2008-99240-116). Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/621760437?accountid=14553

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods, 18*(2), 252. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1661369275?accountid=14553

Carter, N. T., Dalal, D., Zickar, M. J., & Adams, J. E. (2009, April). Do vague quantifiers induce unfolding in personality items? Paper presented at the 24th Annual Meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Carter, N. T., & Zickar, M. J. (2011a). The influence of dimensionality on parameter estimation accuracy in the generalized graded unfolding model. *Educational and Psychological Measurement, 71*(5),

765-788. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/904017569?accountid=14553

Carter, N. T., & Zickar, M. J. (2011b). A comparison of the LR and DFIT frameworks of differential functioning applied to the generalized graded unfolding model. *Applied Psychological Measurement, 35*(8), 623-642. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/920225667?accountid=14553

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*(4), 523-562. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/62286753?accountid=14553

Chernyshenko, O. S. (2003). *Applications of ideal point approaches to scale construction and scoring in personality measurement: The development of a six-faceted measure of conscientiousness* Available from PsycINFO. (620233080; 2003-95010-007). Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/620233080?accountid=14553

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*(1), 88-106. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/621651554?accountid=14553

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1289808194?accountid=14553

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/614317877?accountid=14553

Conn, S. & Rieke, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual.* Champaign, IL: Institute for Personality and Ability Testing.

Coombs, C. H. (1964). A theory of data. New York: Wiley.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*(1), 134-135. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/616876597?accountid=14553

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*(4), 662-680. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/614284143?accountid=14553

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology. (2nd ed.)* (pp. 577-636) Consulting Psychologists Press, Palo Alto, CA. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618385638?accountid=14553

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*(4), 465-476. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/822369070?accountid=14553

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. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618336031?accountid=14553

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*(1), 26-42. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618123922?accountid=14553

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84-96. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/621074376?accountid=14553

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage Publications, Inc, Thousand Oaks, CA. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618050327?accountid=14553

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*(1), 100-112. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.88.1.100

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3-4), 117-144. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618268584?accountid=14553

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. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619142398?accountid=14553

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology, 85*(6), 869-879. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.85.6.869

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology,87*(4), 765-780. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619888820?accountid=14553

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*(3), 530-541. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619737885?accountid=14553

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

Kosinski, M. (2009). Application of the dominance and ideal point IRT models to the extraversion scale from the IPIP Big Five Personality Questionnaire. (Mphil Dissertation) Cam- bridge University.

Retrieved from http://mypersonality.org/wiki/lib/exe/fetch.php?media=mkosinski_irt_2009.pdf

LaPalme, M. L., Wang, W., Joseph, D. L., Saklofske, D. H., & Yan, G. (2016). Measurement equivalence of the Wong and Law Emotional Intelligence Scale across cultures: An item response theory approach. *Personality and Individual Differences, 90*, 190-198. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1803818541?accountid=14553

Le, K., Donnellan, M. B., Spilman, S. K., Garcia, O. P., & Conger, R. (2014). Workers behaving badly: Associations between adolescent reports of the big five and counterproductive work behaviors in adulthood. *Personality and Individual Differences, 61-62*, 7-12. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1016/j.paid.2013.12.016

Levine, M. V. (1984). *An introduction to multilinear formula score theory.* (Personnel and Training Research Programs, Office of Naval Research, Measurement Series No. 84- 4). Arlington, VA: Personnel and Training Research Programs.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22 140*, 55. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/615002361?accountid=14553

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*(4), 251-265. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/61875792?accountid=14553

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. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619374904?accountid=14553

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*(6), 1524-1536. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/621643523?accountid=14553

Nye, C. D. (2011). *The development and validation of effect size measures for IRT and CFA studies of measurement equivalence.* Available from PsycINFO. (1269433733; 2012-99220-321). Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1269433733?accountid=14553

O'Brien, E., & LaHuis, D. M. (2011). Do applicants and incumbents respond to personality items similarly? A comparison of dominance and ideal point response models. *International Journal of Selection and Assessment, 19*(2), 109-118. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/887942785?accountid=14553

Raju, N. S., van, d. L., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement,19*(4), 353-368. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618796438?accountid=14553

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619734866?accountid=14553

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/63718816?accountid=14553

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*(2), 211-233. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619419377?accountid=14553

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3-32. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619524937?accountid=14553

Salgado, J. F. (1997). The five factor model of personality and job performance in the european community. *Journal of Applied Psychology, 82*(1), 30-43. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/0021-9010.82.1.30

Salgado, J. (2002). The big five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment, 10*(1-2), 117-125. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619877040?accountid=14553

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar big-five markers. *Journal of Personality Assessment, 63*(3), 506-516. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618625210?accountid=14553

Speer, A. B., Robie, C., & Christiansen, N. D. (2016). Effects of item type and estimation method on the accuracy of estimated personality trait scores: Polytomous item response theory models versus summated scoring. *Personality and Individual Differences, 102*, 41-45. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1824558619?accountid=14553

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006a). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/621548401?accountid=14553

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006b). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*(1), 25-39. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/621079208?accountid=14553

Stark, S. (2007). MODFIT: Plot theoretical item response functions and examine the fit of dichotomous or polytomous IRT models to response data. Champaign, IL.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-90. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/215042771?accountid=14553

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model-data fit of ideal point and dominance models. *Applied Psychological Measurement, 35*(4), 280-295. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/870285230?accountid=14553

Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. *Organizational Research Methods, 15*(3), 363-384. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1041004283?accountid=14553

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology,44*(4), 703-742. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/618107386?accountid=14553

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]. Skokie, IL: Scientific Software International.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review,34*(4), 273-286. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/614934623?accountid=14553

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology,33*, 529-554. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1086/214483

Thurstone, L. L., & Chave, E. J. (1929). Theory of attitude measurement. *The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church.* (pp. 1-21) University of Chicago Press, Chicago, IL. doi:http://dx.doi.org.proxy2.library.illinois.edu/10.1037/11574-001

Valbuena, N. (2004). *An empirical comparison of measurement equivalence methods based on confirmatory factor analysis (with mean and covariance structures analysis) and item response theory.* Available from PsycINFO. (620630932; 2004-99020-128). Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/620630932?accountid=14553

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619672309?accountid=14553

Van Schuur, Wijbrandt H.; Kiers, Henk A.. (1994). Why Factor Analysis Often is the Incorrect Model for Analyzing Bipolar Concepts, and What Model to Use Instead. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/120012.

Wang, W. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*(3), 221-261. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/62072210?accountid=14553

Wang, W. (2013). *A Bayesian Markov chain Monte Carlo approach to the generalized graded unfolding model estimation: The future of non-cognitive measurement.* Available from PsycINFO. (1676371094; 2015-99080-541). Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1676371094?accountid=14553

Wang, W., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement, 37*(4), 316-335. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/1364721134?accountid=14553

Wong, C., & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. *The Leadership Quarterly, 13*(3), 243-274. Retrieved from http://search.proquest.com.proxy2.library.illinois.edu/docview/619915079?accountid=14553