**CHAPTER 3**

**RESULTS**

**MODEL FIT**

We obtained model-data fit for the GGUM first. To estimate GGUM item parameters, which is required by the Modfit software, no reverse coding was needed. Then 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 on which items to be reversely coded in preparation for applying the dominance models. If as the latent trait level went up, the probability went down of the participants agreeing with the item, then the item was considered a negative item, and thus was reversed.

***The Well-being scale.*** Based on the ICCs given by Modfit, 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 showed 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-data fit was then obtained using theses 17 items for both GGUM and SGR, with negative items reversed for the latter. Results of model fit 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). Therefore, both GGUM and SGR showed good fit for item singles, but some extent of misfit for item doubles and triple. In the current study, we focus on 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, and that item doubles and triples have been found to be sensitive to local independence (Drasgow et al., 1995).

For a 17-item scale measuring a specific personality facet, local dependence is not rare (Chernyshenko et al., 2007), and thus 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 both groups, GGUM had better fit than SGR for item doubles and triples. For item singles, GGUM fitted better than SGR for 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. In the U.S. group, SGR fitted only faintly better than GGUM, while in the Chinese group, GGUM fitted better than SGR.

Considering that generally, polytomous GGUM had slightly better model-data fit than SGR, and that both models showed acceptable, if not very 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 were what one would expect from an item that was working as an unfolding/intermediate item. Another characteristic of an unfolding item is that it probabaly won’t be modeled successfully 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 mode 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, model fit of SGR now became almost as good as GGUM, majorly due to the significant improvement in the 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. Model-data fit with 18 items for both GGUM and 2PLM was computed, which can be found in Table 3. As shown in Table 3, both GGUM and SGR exhibited much better fit than with polytomous data. All combinations of group, model, and item types demonstrated adequate fit except for item triples for the U.S. group under 2PLM, which showed only slight misfit. Same as when with polytomous data, GGUM fitted 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, model fit of 2PLM for 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” for them, which is a situation that GGUM2004 couldn’t deal with without collapsing the responses. But we couldn’t collapse responses, 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 theses 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 Chinses 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, but the misfit was not too severe. Compared with SGR, GGUM showed worse fit in the U.S. group, but better fit in the Chinese group.

Given the fact that the data-model fit was not too bad, we decided to included 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, majorly because 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 adequate fit. The GGUM fitted only faintly better than 2PLM. Item 19 was identified as an unfolding item in both groups under GGUM (Figures 14-15), 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, all model-data fit got worse only slightly. GGUM now fitted slightly worse than 2PLM for the U.S. group, but still moderately better than 2PLM for the Chinese group (Table 8).

**DIF**

***The Well-being scale.*** Under the constrained baseline approach with the 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 identified at least one item as the linking item for the free baseline analysis. However, during the free baseline analysis, many of the matrices again turned out to have been too ill-conditioned to produced correct 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 ≤ *d* < .8; Items 5 and 7), only 2 items exhibited large DIF (i.e., 0.8 ≤ *d*; Items 10 and 15), and the remaining 9 items showed small DIF (.2 ≤ *d* < .5).

***The Curiosity scale.*** With the Curiosity scale, matrices were also ill-conditioned under both baseline approaches, and 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.