**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 one unfolding item showing properties of a good non-monotonic item, GGUM showed equally good or better fit than SGR across 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 model fit of SGR. Although model fit of GGUM improved slightly when the unfolding item was removed, the extend 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. Whether an item was non-monotonic was based on expert ratings. However, since no ICCs and 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 that for the 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.

Due to ill-conditioned matrices, we were not able to carry out DIF anlaysis with constrained and 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 GGUM, which led to the authors deleting those data sets from the simulation (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 output looks all normal. We suggest that researchers supervise the whole process, and that 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. GGUM DIF analyses missing 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, 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. Therefore, the current study contributes to the measurement equivalence and cross-cultural personality literature. Also, contrast 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).