**Value personality as an IOer – explain**

Two major issues when personality tests are used – if IRT, model selection, which concerns item writing – majorly the inclusion of intermediate items; -- ME, which concerns fairness – important as today hiring has become more and more international

**Master’s thesis:** covered both – IRT model selection, and DIF analyses – SGR and GGUM fitted equally well, GGUM slightly better – effects of intermediate items – working, not working

DIF: GGUM2004 using MML may not be the most ideal method – Bayesian?

Effect sizes showed that NHST results may be too sensitive to large samples – most items having small to moderate DIF according the effect sizes, though they are all flagged as “significant” by NHST

Future: in my dissertation, I will explore by experiment response processes of participants to different types of intermediate items (FACT) – match between researchers’ assumptions about response processes and the reality; including reference for comparison

**The PISA project:** ME with CFA – partial invariance model; SEM predicting math achievement; ecological fallacy

IRT and FA&SEM – personality data

Scale development

I am currently a fourth-year PhD student in Industrial-Organizational Psychology at the University of Illinois at Urbana-Champaign with an interest in contributing to a deeper understanding of the IRT modelling and measurement equivalence of personality data

Puzzle: IRT based modeling – Master’s thesis – DIF and model selection; Personality assessment ME – PISA; Personality scale development – CCS revision.

**The paragraph about my Master’s thesis:  
Current work**

Personality, besides cognitive ability, has been recognized as one of the most important predictors of a variety of work-related outcomes, including turnover, task performance, organizational citizenship behavior (OCB), counterproductive work behavior (CWB), and job satisfaction. Today, hiring is no longer limited to a single country or culture, but has become more and more global, hence the importance of applying a culturally equivalent personality test on such occasions. When an IRT approach is adopted for analyzing personality data, model selection has become a major issue. The debate has long been over whether a dominance IRT model or an ideal point IRT model is more appropriate for such analyses. 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. Whereas the ideal point methods, inspired by a series of Thurstone’s studies on measuring attitudes, hypothesize that the closer the statement is to a respondent’s trait level, the higher the probability of endorsement.

Therefore, my Master’s thesis adopted an IRT approach to studying differential item functioning (DIF) of the Well-being scale and the Curiosity scale from the Comprehensive Personality Scale (CPS), using data collected from China and the United States. In my study, in an NHST paradigm, I applied to the data a dominance IRT model, the Samejima’s Graded Response Model (SGR), and an ideal point model, the Generalized Graded Unfolding Model (GGUM), and compared their performance in modeling the data, as well as detecting DIF. A DIF effect size measure was also adopted in order to obtain the magnitude of DIF, and thus compensate for the oversensitivity of NHST to large samples. How well the two IRT models describe the data was examined via model-data fit, and DIF analyses were carried out by first figuring out the linking items through the constrained baseline approach, and then flag DIF items using the free baseline approach. DIF was indexed using a log-likelihood ratio (LR) test.

It was found that in terms of describing the personality data, GGUM in general had moderately better fit than SGR, when as few as only one item turned out working as a good intermediate item, i.e., having a non-monotonic, or unfolding, instead of a monotonic item characteristic curve (ICC) that goes up along the theta axis. The removal of such intermediate items, on the other hand, 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. The existence of working unfolding items is likely to harm the fit for SGR, but not or not as much as for GGUM. These findings are important because first, they support the argument that GGUM is a more flexible model than SGR for fitting personality data, as besides extreme items, it’s also able to model intermediate items with non-monotonic ICCs, which SGR simply can’t handle. Second, they have put a spotlight on the “true” intermediate items in the course of selecting the appropriate IRT model to analyze response data. The “true” intermediate items are items that turn out to have bell-shaped ICCs based on the data, rather than the ones that are only rated as intermediate in terms of extremity by researchers in a lab.

In terms of DIF, almost all items were flagged as DIF items under SGR, while according to the DIF effect sizes, many of them only had moderate to medium DIF. **As applicable as SGR is for DIF analyses of cross-cultural personality tests, the effect sizes of DIF should also be taken into consideration, in order to reach a more accurate perception of the severity of DIF.**

DIF couldn’t be examined for GGUM due to ill-conditioned matrices produced by GGUM2004 using the marginal maximum likelihood (MML) technique, but test characteristic curves (TCCs) were used in the study to examine test differential functioning (TDF). It was found that both scales exhibited smaller DTF under GGUM than SGR, which means according to GGUM, both the Well-being and the Curiosity scales are more equivalent cross cultures than when they are examined with SGR. These findings raised questions for the appropriateness of the widely used MML technique for GGUM analyses, and thus calls for the application of alternative techniques such as the Bayesian estimation. The finding that both scales were shown to have smaller TDF under GGUM than under SGR showcased GGUM’s potential to be adopted in DIF analyses for personality tests, considering that in modeling personality data, GGUM is more flexible, and does an even better job than SGR.

This thesis is the first to illustrate that GGUM is a promising IRT model to be applied to analyzing DIF for cross cultural personality tests.