Doi: https://doi.org/10.1027/1015-5759/a000540

ESM 2.

Explanation of the exact procedure of testing for MI using moderated factor analysis

The traditional method to establish MI and latent mean differences across gender and age is multi-group CFA in which increasingly restrictive models are fit to evaluate MI. That is, three levels of MI are evaluated in a hierarchical order (Byrne & van de Vijver, 2010): configural, metric, and scalar MI. MI testing starts with establishing the *configural invariance* model, assuming equal factor structure across groups. In the next step, *metric invariance* (equal factor loadings) was evaluated to investigate whether the strength of the relationship between each item and its factor is the same for all groups. Finally, *scalar invariance* was tested, holding item intercepts equal across groups. Scalar invariance is used to explore whether individuals with the same values on the latent construct would have equal values on the observed variables. If metric invariance is established, structural relationships can be meaningfully compared across groups, whereas the comparison of group means is meaningful only if scalar invariance is confirmed (Schmitt, Golubovich, & Leong, 2011).

As we wanted to establish MI with respect to a continuous variable (age), we used the method of moderated factor analysis (MFA) (Bauer, 2017; Molenaar, Dolan, Wicherts, & Van der Maas, 2011) instead of the traditional procedure above. Advantage of MFA over multi-group CFA is that 1) the age variable does not need to be categorized so that its continuous nature can be retained in the MI analysis, and 2) MI can be tested more easily with respect to gender and age simultaneously such that possible interactions between age and gender in MI can be

accounted for (e.g., factor loadings may be increasing across age, but with a different rate for males as compared to females).

In MFA, the parameters of the factor model (in our case: the loadings the threshold parameters, the factor means, and the factor variances) are moderated by the variables of interest (in our case: gender and age). For instance, if the loadings and threshold parameters are moderated, they are allowed to differ across ages and across gender (which, in a more traditional multi-group approach corresponds to allowing the loadings and thresholds to differ freely across gender and age groups). The notion of MI requires that loadings and thresholds parameters are the same across gender and across age, while differences at the latent level (i.e., differences on the PA factor and NA factor) across gender and age are allowed. Therefore, to test for MI using MFA, it is tested in a step-wise fashion whether a model with moderation of the latent mean and latent variance of the factors (MI model) fits better than a model with moderation of the loadings and the threshold parameters (no-MI model).

Commonly, parameters of the factor model are moderated using a linear function. For gender this is not a problem as it is a dichotomous variable. However, for age, this implies that factor loadings, thresholds, factor means, and factor variances are tested to increase or decrease linearly across age. As age trends may not be strictly linear, we consider both age and quadratic age effects in the MI analysis. As a result, we have the following five moderators: gender, age, age-squared, gender × age, and gender × age-squared.

The exact procedure that we will use is as follows: We start with a Baseline Model (no-MI assumed) which corresponds to the configural model discussed above. In the Baseline Model all loadings and thresholds are moderated by our five moderators. Next, we start dropping moderators from the loadings while introducing the moderator at the latent level (i.e., moderation

of the factor variance). These models correspond to the metric model in the traditional approach above. We first drop the interaction moderators (gender × age, and gender × age-squared), then we drop the age moderators (age and age-squared), and finally we drop the gender moderator. Note that every time we drop a moderator from the loadings, we introduce its effect at the factor variance to account for differences at the latent level across that moderator (which is allowed in the case of MI). After all moderators are dropped for the loadings and introduced in the variances, we start dropping moderation of the thresholds. These models correspond to the scalar model in the traditional procedure above. Similar as for the factor loadings, we first drop the interaction moderators (gender × age, and gender × age-squared), then the age moderators (age and age-squared), and then the gender moderator. Each time we drop a moderator from the thresholds, we introduce its effect at the factor mean to account for latent mean differences across the moderator (which is again allowed in the case of MI). The final model (full MI-model) will thus consist of moderation of the factor mean and factor variance by our five moderators, but no moderation of the loadings and the thresholds. If MI holds, this final, full-MI model is the best fitting model.

Byrne, B.M., & van de Vijver, F.J.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*, 107-132.

Schmitt, N., Golubovich, J., & Leong, F.T. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using Big Five and RIASEC measures. *Assessment*, *18*, 412-427.