

# Changing the reference group in multiple-group scalar invariance factor analysis SEM and ESEM models

December 1, 2017

Consider a multiple-group scalar invariance factor model

$$Y_{ig} = \nu + \Lambda\eta_{ig} + \varepsilon_{ig}$$

$$\eta_{ig} \sim N(\alpha_g, \Psi_g)$$

where  $\eta_{ig}$  and  $Y_{ig}$  are the factors and the factor indicators for observation  $i$  in group  $g$ . The loading matrix  $\Lambda$  contains only free parameters and the scale of the factors is set by assuming that the factor means  $\alpha_1$  in the first group are 0 and that the factor variances in the first group are 1, i.e.,  $\Psi_1$  is a correlation matrix with 1 on the diagonal. The first group is called the reference group. If we change the reference group, i.e., we rearrange the data so a different group appears as the first group the log-likelihood of the model will remain the same however the parameter estimates will change as follows. If the new reference group is group  $k$  the new factor means and variances for group  $g$  are

$$\eta_{ig} \sim N((\alpha_g - \alpha_k)C_k, C_k\Psi_g C_k^T)$$

where  $C_k$  is the vector  $1/\sqrt{\text{diag}(\Psi_k)}$

If one is interested in testing for significant difference in factor means between groups  $g_1$  and  $g_2$ , changing the reference group will change the tested hypothesis from

$$0 = \alpha_{g_1} - \alpha_{g_2}$$

to

$$0 = (\alpha_{g_1} - \alpha_{g_2})C_k.$$

While these two hypotheses are logically identical and are guaranteed to asymptotically have identical conclusion (i.e. we would conclude that the means are significantly different or not as long as we have sufficiently large sample size regardless of which one of the two hypotheses we use), in finite sample size the two hypotheses will not have identical p-values. When using maximum-likelihood estimation, the testing of the above hypotheses is based on the delta method and an asymptotic computation. When a hypothesis reparameterization is not a linear transformation and involves multiplication of model parameters (such as in the above example) the p-value is not preserved (the simplest example is the fact that testing if a parameters  $a$  is significant we get twice as large T-test statistic as compared to testing  $a^2$ ). The Bayesian estimator has an advantage over ML estimator because it preserves the  $p$ -value in this situation of testing significant differences between the factor means and generally behaves more intuitively when it comes to model reparameterizations.

In our experience, even though the p-value changes when the reference group is changed, in most cases the conclusion does not change (i.e., regardless of which reference group we use the conclusion remains the same). Rarely though the conclusion will indeed change, and perhaps using the Bayesian estimator as an alternative is the sound way to resolve this issue in such cases. The above discussion applies both for SEM and for ESEM models.

In Mplus it is easy to study this issue further because of the flexibility of model constraints. Consider this two-group example

model: f1-f2 by y1-y6 (\*1);

model g2: f1-f2 (v1-v2); [f1-f2] (m1-m2);

model constraints: new(a1-a2); a1=m1/sqrt(v1); a2=m2/sqrt(v2);

The statistical significance of a1 and a2 will be the same as the one of m1 and m2 if the two groups are reversed.