Methods Draft

Jimmy

Methods Draft

Method

Simulation Design

Adapted from Hsiao et al. (2021), the current simulation study aimed to compare performance of UPI and RAPI with that of 2S-PA-Int on estimating latent interaction effects for continuous congeneric items. We investigated the bias and variance of interaction estimates generated by the three methods over various levels of sample size, reliability, and correlation between first-order latent variables. The generated population data was based on the model below with predefined parameter values:

$$\begin{aligned} x_i &= \tau_{x_i} + \lambda_{x_i}\xi_x + \delta_{x_i}; \\ m_i &= \tau_{m_i} + \lambda_{m_i}\xi_m + \delta_{m_i}; \\ y &= \tau_y + \gamma_x\xi_x + \gamma_m\xi_m + \gamma_{xm}\xi_x\xi_m + \zeta, \end{aligned} \tag{1}$$

where the path coefficients of two latent predictors (i.e., γ_x and γ_m) and their interaction term (i.e., γ_{xm}) were all set to 0.3 for the structural model. The first-order latent predictors ξ_x and ξ_m were simulated from standard normal distributions with means of 0 and variances fixed at 1, each indicated by three items (i.e., ξ_x indicated by [x_1, x_2, x_3]; ξ_m indicated by [m_1, m_2, m_3]). The first-order indicators and the dependent variable y were all observed continuous variables with normally distributed error. Accordingly, δ_{x_i}, δ_{m_i} and ζ were assumed to have multivariate normal distributions and be mutually independent. τ_{x_i}, τ_{m_i}, and τ_y were their corresponding constant intercepts and assumed to be 0. The first-order indicators were mean-centered for UPI, RAPI and 2S-PA-Int at the sample level.

Drawing from Jöreskog's (1971) concept, congeneric tests were defined as a set of observed items measuring a latent construct with different factor loadings and unique error terms. These error terms were assumed to be uncorrelated with each other and with the latent construct, reflecting random measurement error unique to each item. To align with

this concept, we manipulated the factor loadings and error variances of first-order indicators to create sets of congeneric items in the measurement model. Specifically, the first, second, and third indicators were set to fixed values of 1.0, 0.9, and 0.75 for both first-order latent variables (i.e., $\lambda_{x_1} = \lambda_{m_1} = 1.0$, $\lambda_{x_2} = \lambda_{m_2} = 0.9$, $\lambda_{x_3} = \lambda_{m_3} = 0.75$). We involved reliability estimates to manipulate error variances since equation (9) demonstrates that the error variance of the interaction term was a function of first-order indicators' reliability, implying that the interaction effect could be impacted by the amount of measurement error. Hence we included reliability as a varying condition to explore how each method performed under three reliability conditions: .70, .80, and .90, which resulted in three levels of error variances. For each level of error variance, we systematically manipulated proportions of error variances each first-order indicator occupied. The proportions were set to maintain consistency with the design in Hsiao et al. (2021): 44% of the total error variance for the first indicator, 33% for the second, and 23% for the third. Then we obtained the manipulated error variances according to equation (8). For instance, $\theta_{x1},\ \theta_{x2},\ \theta_{x3}$ and $\theta_{m1},\ \theta_{m2},\ \theta_{m3}$ were $[3.01,\ 1.76,\ 0.78]$ when $\lambda_{x1},\ \lambda_{x2},\ \lambda_{x3} = \lambda_{m1},\ \lambda_{m2},\ \lambda_{m3} = [1,\ 0.9,\ 0.75]$, as the reliability was varied at .70, .80, and .90 respectively.

Following the recommendation by Marsh et al. (2004), $\xi_x\xi_m$ was represented through a matched-pair configuration of indicators in the UPI method, namely x_1m_1, x_2m_2, and x_3m_3. For the RAPI and 2SPA methods, $\xi_x\xi_m$ was loaded by single PIs. Specifically, for RAPI the interaction term's PI was the mean scores of first-order indicators, while for 2S-PA-Int was pre-computed Bartlett factor scores. To reduce the problem of multicollinearity between first-order latent predictors and the interaction term, the DMC strategy was applied to all the methods.

The literature on latent interaction methods showed a range of researcher-selected sample sizes from 20 to 5,000 (Chin, Marcolin, & Newsted, 2003; Lin et al., 2010; Cham et al., 2012), with common selections ranging from 100 to 500. Consequently, we selected N =

100, 250, and 500 to represent small, medium, and large sample sizes, respectively.

Based on the study design in Hsiao et al. (2021), we pre-specified three population correlations between latent predictors ($Corr[\xi_x, \xi_m]$): 0, 0.3, 0.6 as zero to large correlation. Given that the variance of y (i.e., σ_y^2), $\sigma_{\xi_x}^2$, and $\sigma_{\xi_x}^2$ was set to 1, ψ could be computed as $1 - R^2$ in which $R^2 = \gamma_x^2 + \gamma_m^2 + 2\gamma_x\gamma_m Corr[\xi_x, \xi_m] + \gamma_{xm}^2(1 + Corr[\xi_x, \xi_m]^2)$. Take $Corr[\xi_x, \xi_m] = 0$ as an example,
$\psi = 1 - (0.3^2 + 0.3^2 + 2 \times 0.3 \times 0.3 \times 0 + 0.3^2 \times (1 + 0)^2) = 0.73$. Similarly, ψ was determined to be 0.668 and 0.590 for $Corr[\xi_x, \xi_m]$ equal to 0.3 and 0.6, respectively.

In summary, our study implemented a $3 \times 3 \times 3$ factorial design, accommodating variations across three sample sizes, three levels of correlation between first-order latent predictors, and three levels of reliability.

Evaluation Criteria

We chose widely used evaluation criteria that were summarized from 2000 replications to assess the accuracy and precision of interaction effect estimates (γ_{xm}) of the three methods.

Averaged Raw Bias and Standardized Bias

Standardized bias (SB) was used to evaluate averaged raw bias and accuracy of parameter estimates. It provided a normalized measure that allowed for comparing bias across different scales or units of measurement, and reflected how far an estimate was from its true value in standard error units. Hence SB was useful in comparisons where models often contained a variety of parameter types (e.g., factor loadings, path coefficients).

The Standardized Bias (SB) was defined through the averaged raw Bias (B):

$$SB = \frac{B(\gamma_{xm})}{SE_{\gamma_{xm}}}, \tag{2}$$

$$B(\gamma_{xm}) = R^{-1}\Sigma_{r=1}^{R}(\hat{\gamma}_{xm_r} - \gamma_{xm}), \tag{3}$$

where R was the total number of replication cycles that were counted from 1 to 2,000. $\hat{\gamma}_{xm_r}$ was the estimated interaction effect in each replication cycle r and γ_{xm} was the population parameter set at 0.3. $B(\gamma_{xm})$ was the averaged deviation of estimates, $\hat{\gamma}_{xm}$, from 0.3, and $SE_{\gamma_{xm}}$ represented the empirical standard error of $\hat{\gamma}_{xm}$ across replications. Collins et al. (2001) suggested that an absolute value of SB ≤ 0.40 would be considered acceptable for each replication condition.

Coverage Rate

The coverage rate with a 95% confidence interval (CI) served as a critical metric for evaluating the reliability and accuracy of simulation results. It was defined as the percentage of replications in which the Wald confidence interval captured the true interaction effect γ_{xm}. Low coverage rates meant that the proportion of times that γ_{xm} fell within the CI across replications was low, indicating that the model might have issues of misspecification, inappropriate estimation methods, small sample sizes, or violations of statistical assumptions. A coverage rate larger than 91% was considered acceptable (Muthén & Muthén, 2002).

Robust Relative Standard Error Bias and Outlier Proportion of SE

The relative standard error (SE) bias was used to evaluate the precision of $\hat{\gamma}_{xm}$. This criterion compared the empirical standard deviation of $\hat{\gamma}_{xm}$ with the sample-estimated standard error across replications:

$$Relative\ SE\ Bias = \frac{R^{-1}\Sigma_{r=1}^{R}(\widehat{SE}_r - SD)}{SD}, \tag{4}$$

where $\widehat{SE}_r$ was the sample-estimated standard error of $\hat{\gamma}_{xm}$ in a single replication cycle r and SD is the empirical standard deviation obtained from all replications. With SD being used as a reference variability measure of $\hat{\gamma}_{xm}$, smaller relative SE bias meant the

estimated standard errors were closer to the referenced variability, and the uncertainty of $\hat{\gamma}_{xm}$ across replications was more accurately measured in each simulation condition. Absolute values of relative SE bias $\leq 10\%$ were considered acceptable and indicated that the standard errors were reasonably unbiased (Hoogland & Boomsma, 1998). SEM typically required a relatively large sample size to obtain sufficient information to reliably estimate model parameters. Insufficient sample sizes might result in largely biased SEs due to increased uncertainty around the parameter estimates (Bollen & Long, 1993; Byrne, 2016). Given that the conditions of small sample size (N = 100) and high amount of measurement error ($\rho = 0.7$) were included in this study design, a robust version of relative SE bias was calculated as an alternative to the regular one:

$$\textit{Robust Relative SE Bias} = \frac{\widehat{MDN}(SE_r) - MAD}{MAD}, \tag{5}$$

where $\widehat{MDN}(SE_r)$ represented the median value of the estimated SE values and MAD is the empirical median-absolute-deviation of SE values. In the context of biased SEs, we did not assume a specific distribution of SEs (e.g., normal distribution) and hence we used the median due to its robustness to non-normal distributions with skewed data and outliers (Rousseeuw & Hubert, 2011). In addition, MAD measured variability around the median and could serve as a robust alternative to standard deviation that could be inflated by outliers or non-normality (Daszykowski et al., 2007). Besides, an outlier detection using the interquartile range (IQR) method was included as a supplemental information of SE estimates:

$$O_a \notin (Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR), \tag{6}$$

where O_a was an observation of outlier for $a = 1, 2, \ldots,$ b. IQR captured the spread of the middle 50% of the sample SEs by $IQR = Q_3 - Q_1$, where Q_1 and Q_3 were the 25th percentile and the 75th percentile of the sample. The outlier proportion was then calculated by b/R where b represented the total count of identified outliers. Like the robust relative SE bias, the IQR method did not rely on the assumption of normal distribution, thus making it versatile across any distribution.

Root Mean Squre Error

The last criterion was the root mean square error (RMSE), calculated by taking the squared root of the sum of squared bias:

$$RMSE = \sqrt{R^{-1}\Sigma_{r=1}^{R}(\hat{\gamma}_{xm_r} - \gamma_{xm})^2}. \tag{7}$$

It quantified the average magnitude of the difference between the interaction estimates and the true value, reflecting both the bias and variability of the estimates across replications. Under one condition across 2,000 replication, a smaller RMSE value of a method indicated that it had relatively more accuracy than the other two methods in estimating $\hat{\gamma}_{xm}$ (Harwell, 2019). RMSE was most informative when comparing across methods under the same simulated conditions by isolating factors of sample size, model complexity, and the amount of disturbance.