2S-PA-Int-Cat

Gengrui (Jimmy) Zhang[1]

[1] University of Southhern California

Author Note

The authors made the following contributions. Gengrui (Jimmy) Zhang: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Gengrui (Jimmy) Zhang. E-mail: gengruiz@email.com

Abstract

Two-stage path analysis with interaction for categorical variables.

Keywords: keywords

Word count: X

2S-PA-Int-Cat

Methods

We adopted a fully crossed design with varying conditions of sample size, composite reliability of scale, interaction effect, and item skewness, based on the study design of Aytürk et al. (2020) and (Hsiao & Lai, 2021). We compared the accuracy of UPI with three product–indicator formations (all-pair, matched-pair, and parceled-pair), LMS for categorical items, and 2S-PA-Int in recovering the latent interaction effect. A summary table of study design was shown in Table (To be updated).

The simulation script was drafted and structured using SimDesign package; the dataset was generated and analyzed using Mplus 8.8 (Muthén & Muthén, 1998/2017); the entire simulation study was run using MplusAutomation 1.1.0 (Hallquist & Wiley, 2018) on R 4.5.1 (R Core Team, 2025).

Population Structural Model

We considered a latent regression model with two exogenous latent variables for person j with $j = 1, \ldots, N$, ξ_{x_j} and ξ_{j}, and one endogenous latent outcome, ξ_{j}, as our population model. The primary estimand was the standardized latent interaction effect of ξ_{x_j} and ξ_{m_j} on ξ_{y_j}, denoted γ_{xm}:

For $j = 1, \ldots, N$,

$$\begin{bmatrix} \xi_{x_j} \\ \xi_{m_j} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & Corr \\ Corr & 1 \end{bmatrix} \right), \quad \xi_{y_j} = \alpha + \gamma_x \, \xi_{x_j} + \gamma_m \, \xi_{m_j} + \gamma_{xm} \, \xi_{x_j} \xi_{m_j} + \zeta_j, \quad (1)$$

where α was the constant intercept set to 1.2. ξ_{x_j} was the first-order latent predictor with a fixed main effect $\gamma_x = 0.3$, and ξ_{m_j} was the first-order latent moderator also with a

fixed main effect $\gamma_m = 0.3$. Both exogenous factors were factors were standardized with zero means and unit variances. Additionally, they were pre-specified with a fixed correlation of $Corr = 0.3$, and they were allowed to freely correlated with the latent interaction term. We examined two population values of the latent interaction effect: $\gamma_{xm} = 0$ to test the null hypothesis (H_0) and $\gamma_{xm} = 0.3$ (medium effect; Cohen, 1992) to test the alternative hypothesis (H_1). The variance of latent outcome variable ξ_{y_j} was set to 1 when $\gamma_{xm} = 0$ under the H_0 condition, so that the variance of disturbance term $\sigma^2_\zeta = 1 - \Big(\gamma^2_x + \gamma^2_m + \gamma^2_{xm} + 2\,\gamma_x\gamma_m\,Corr\Big)$. Therefore, the values of σ^2_ζ were adjusted according to the latent interaction effect size (e.g., $\sigma^2_\zeta = 1 - (0.3^2 + 0.3^2 + 0 + 2 \times 0.3 \times 0.3 \times 0.3) = 0.766$ when $\gamma_{xm} = 0$, which indicated that the first-order latent predictors and the latent interaction term jointly contributed to explain 23.3% variance in ξ_{y_j}.

Population Measurement Model

After generating person-level latent scores from the structural model, we created observed indicators for each construct. Let i index observed items with $i = 1, \ldots, P$. The latent outcome variable ξ_{y_j} was measured by three continuous indicators ($y_{1j} - y_{3j}$) with differential factor loadings:

$$y_{ij} = \lambda_{y_i}\,\xi_{y_j} + \delta_{y_{ij}}, \qquad i = 1, 2, 3, \tag{2}$$

where the factor loadings λ_{y_i} were unstandardized and fixed at $\{.50,\ .70,\ .90\}$, and the error variances followed normal distribution with $\delta_{y_{ij}} \sim \mathcal{N}\Big(0,\ 1 - \lambda^2_{y_i}\Big)$.

For the predictor ξ_{x_j} and the moderator ξ_{m_j}, we followed Aytürk et al. (2020) and assigned three items to ξ_{x_j} (i.e., $x_{1j}, \ldots, x_{3j}$) and twelve items to ξ_{m_j} (i.e., $m_{1j}, \ldots, m_{12j}$). For each item we drew an underlying continuous precursor using a normal–ogive (cumulative

probit) graded-response specification with item-specific factor loadings (Cho, 2023):

$$x_{ij}^{*} = \lambda_{x_i}\,\xi_{x_j} + \delta_{x_{ij}}, \qquad i = 1, 2, 3, \tag{3}$$

$$m_{ij}^{*} = \lambda_{m_i}\,\xi_{m_j} + \delta_{m_{ij}}, \qquad i = 1, \ldots, 12, \tag{4}$$

where x_{ij}^{*} was the score of the underlying latent continuous variable for each observed categorical item i. Using the normal-ogive metric, factor loadings for ξ_{x_j} were fixed at $1.7 \times \{0.60, 0.70, 0.80\}$, monotonically increasing across the three items, whereas the ξ_{m_j} loadings were fixed at $1.7 \times \{0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85\}$ across the twelve items, reflecting a broad range of discrimination parameters. All factor loadings were unstandardized. The individual-specific error term for each observed indicator i (e.g., $\delta_{x_{ij}}$) followed a normal distribution with a zero mean and a variance $\theta_{x_{ij}}$. To ensure the items had different factor loadings and different error variances while keeping overall measurement quality fixed, we adjusted each item's error variance before categorizing responses. Specifically, we computed the total error variance implied by the target composite reliability and distributed that total across items using a decreasing weight vector from 0.8 to 0.2 (i.e., renormalized to sum to one), which yielded heterogeneous error variances. The manipulation should achive congeneric measurement while preserving the intended composite reliability. The detail was explained in the subsection of reliability.

Throughout, ξ_{x_j}, ξ_{m_j}, the item error terms (δs), and the distrubance term (ζ) were assumed jointly multivariate normal with mean zero and mutually uncorrelated.

Categories and Symmetry of Observed Indicators

We mapped the continuous precursors (i.e., x_{ij}^{*} and m_{ij}^{*}) to observed categorical responses using item-invariant thresholds within each construct. For a generic item,

categories were assigned according to

$$x_{ij} = \begin{cases} 0 & \text{if } x^*_{ij} < \beta_{x_{i1}} \\ k & \text{if } \beta_{x_{ik}} \leq x^*_{ij} < \beta_{x_{i(k+1)}} \\ K-1 & \text{if } \beta_{x_{i(K-1)}} \leq x^*_{ij} \end{cases}, \tag{5}$$

where β_{ik} was the threshold parameter between the kth and $(k+1)$th category for $k = 1, 2, ..., K$.

Specifically, we assigned category 0 if the continuous precursor was below the first threshold, assigned category k and if it fell between the kth and $(k+1)$th thresholds, and assigned the top category if it was above the last threshold. In our design, the items of ξ_{x_j} were all binary, so the rule reduced to a single cut-point: $x_{ij} = 0$ if x^*_{ij} was below the threshold and $x_{ij} = 1$ otherwise. In the symmetric condition, the threshold was set to 0, so that the proportions of 0 and 1 were both 50%. For skewed distribution, the threshold was fixed to 0.9, producing a positively skewed distribution with fewer 1s than 0s.

We applied the same rule to items of ξ_{m_j}, which were designed to have five categories. Here, four common thresholds define the five categories. We used $(-1.5, -0.5, 0.5, 1.5)$ for the symmetric condition and $(0.05, 0.75, 1.55, 2.55)$ for the skewed condition, which were suggested by Aytürk et al. (2020). The exact proportions might differ slightly by item due to congeneric property, but these threshold sets reliably created the intended symmetric versus positively skewed patterns while keeping the cut-points the same for all twelve items.

Sample Size

We varied the total sample size across four levels, $N \in \{100, 250, 500, 2000\}$. The $N = 100$ condition represents a deliberately demanding small-sample setting for detecting interaction effects, which are known to have comparatively low statistical power in field designs (McClelland & Judd, 1993). The $N = 250$ condition places the design just above the range where estimation with ordinal indicators typically begins to stabilize relative to very

small samples. Prior simulation work on confirmatory factor analysis (CFA) or structural equation modeling (SEM) with ordinal data showed that small N could yield bias or instability, with performance generally improving as N moved into the low hundreds (Flora & Curran, 2004; Li, 2015) The $N = 500$ condition reflected a common medium-to-large sample in applied SEM and provided substantially greater power for the interaction than smaller samples. Finally, $N = 2000$ approximated a large-sample regime intended to probe near-asymptotic behavior of the estimators. Related interaction simulations often juxtaposed moderate samples (around $N \approx 200$) with very large samples to benchmark asymptotics (Aytürk et al., 2020; Cham et al., 2013).

Reliability

As we mentioned before, we varied error variances of observed indicators for ξ_{x_j} and ξ_{m_j} according to three reliability levels: $\rho \in \{0.7, 0.8, 0.9\}$. For congeneric items with unit variances, the composite reliability could be computed as: $\rho = \Sigma(\lambda_i)^2/(\Sigma[\lambda_i]^2 + \theta_i)$ (Raykov, 1997; McDonald, 1999). To reach a chosen target reliability, we computed the total residual variance required, $\Theta_{total} = \left(\Sigma[\lambda_i]\right)^2 \frac{1-\rho}{\rho}$, and then allocated this total across items using a decreasing weight pattern (i.e., from 0.8 to 0.2, and rescaled to sum to one). The manipulation set each item's residual variance to $\theta_i = w_i\,\Theta_{\text{total}}$ where w_i represented the item weight. For instance, for the three ξ_{x_j} items with $\Sigma(\lambda_i) = 3.57$ and $\rho = 0.8$, the required Θ_{total} should be 3.19. Using our provisional weights (0.8, 0.5, 0.2) with rescaling would generate $\boldsymbol{w_{x_j}} = (0.53, 0.33, 0.13)$, and hence $\boldsymbol{\theta_{x_j}} = (1.70, 1.06, 0.43)$.

Condidate Analysis Models

We conducted a Monte Carlo simulation study to examine our proposed method, 2S-PA-Int, of estimating latent interaction effects, and compare with a few widely used approaches.

2S-PA-Int. The 2S-PA-Int approach separated measurement from structural estimation while correcting for measurement error in the predictors and their interaction. In

Stage 1, we estimated person-specific factor scores and posterior standard errors (SEs) for the exogenous constructs and the outcome. The three binary x_j indicators were fit with a 2-parameter logistic (2PL) IRT model, and the twelve ordered m_j indicators with a graded response model (GRM), both using the `mirt` package. From these models we obtained expected-a-posteriori (EAP) factor scores $\hat{\xi}_{x_j}$ and $\hat{\xi}_{m_j}$ and their SEs. For the continuous y_{ij} indicators, we estimated a single-factor model with unit-variance scaling and extracted the factor scores $\hat{\xi}_{yj}$ with SEs. We again adopted the double-mean-centered strategy by mean centering factor score items $\hat{\xi}_{x_j}$ and $\hat{\xi}_{m_j}$ and then formed a mean-centered product factor score $\widehat{\xi}_{xm_j} = \widehat{\xi}_{x_j}\widehat{\xi}_{m_j} - \overline{\widehat{\xi}_{x_j}\widehat{\xi}_{m_j}}$ for ξ_{xm_j}. In Stage 2, we treated each score as a single indicator for the corresponding latent variables in Mplus and passed case-specific constants via definition variables. For each latent variable, we fixed the loading of the single indicators to their reliability proxy and fixed the residuals to the implied error variances (see Introduction).

UPI methods. We estimated three UPI specifications that differ only in how product indicators for the interaction term ξ_{xm_j} were constructed from the observed indicators. In every case, observed items of ξ_{x_j} and ξ_{m_j} were mean-centered prior to PI formation, and the product variables were further double mean–centered to reduce multicollinearity between formed PIs and firsr-order indicators (Lin et al., 2010). For all-pair UPI, we created all $3 \times 12 = 36$ PIs by crossing each centered x_j item with each centered m_j item. In the matched-pair UPI model, we ranked items by computing their item reliabilities, selected the three most reliable m_j items, and formed three one-to-one products with all x_j items. An example of computing item reliability was $\rho_{x_1} = \lambda^2_{x_1}/(\lambda^2_{x_1} + \theta_{x_1})$. Last, since m_j had more indicators than x_j items we parceled m_j items and then formed PIs by pairing each parcel with one x_j item. Concretely, we adopted the forming strategy suggested by Rogers and Schmitt (2004) that mixed high-, mid-, and low-loading items to create item parcels, and then took averages of the three parcels. Consequently, all m_j items finally formed three parcels, such as $P_{m_1} = (m_1, m_{12}, m_4, m_9)/4$, $P_{m_2} = (\{m_2, m_{11}, m_5, m_8)/4$, and

$P_{m_3} = (\{m_3, m_{10}, m_6, m_7)/4$. Then we ranked item reliability of x_i items and paired them to parcels (i.e., in our case, x_1 was paired with P_{m_1}, x_2 was paired with P_{m_2}, x_3 was paired with P_{m_3}).

LMS-Cat. The LMS-cat model was estimated by explicitly treating ξ_x and $\xi_m$items as ordered categorical in Mplus script (i.e., $CATEGORICAL = x_1\ x_2\ ...$). The latent interaction was estimated via maximum likelihood with numerical integration using Mplus's default STANDARD quadrature (adaptive rectangular) with 15 integration points per dimension.

Evaluation Criteria

For each method, we computed convergence rate, standardized bias, relative standard error (SE) bias, root mean squared error (RMSE), empirical Type I error, and empirical statistical power, to compare these indices to examine the performance of each method on estimating the latent interaction effect.

Convergence Rate. For each replication, the program may or may not produce an error, such as non-positive definite variance-covariance matrix or negative variance estimates, depending on the random simulated sample. The convergence rate was calculated as the proportion of replications that did not generate any error messages out of all replications. Sometimes extreme parameter values and standard errors could appear especially in small sample size (i.e., $N = 100$) even though no error messages are generated. Thus, robust versions of bias, relative SE bias, and RMSE values were used.

Standardized Bias. The standardized bias was used to evaluate how far an estimate is from its true value in standard error units. It was defined using the raw bias and standard error of a point estimate:

$$B(\gamma_{xm}) = R^{-1}\Sigma_{r=1}^{R}(\hat{\gamma}_{xm_r} - \gamma_{xm}), \tag{6}$$

$$SB = \frac{B(\gamma_{xm})}{SE_{\gamma_{xm}}}, \tag{7}$$

where R was the total number of replications for $r = 1, 2, \ldots, 2{,}000$. $B(\hat{\gamma}_{xm})$ was the averaged deviation $\hat{\gamma}_{xm}$ from the population parameter γ_{xm}, and $SE_{\hat{\gamma}_{xm}}$ was the empirical standard error of $\hat{\gamma}_{xm}$ across replications. An absolute value of $SB \leq 0.40$ was considered acceptable for each replication condition (**collinsComparisonInclusiveRestrictive2001?**).

Robust Relative Standard Error (SE) Bias. The robust relative SE bias was computed as:

$$\textit{Robust Relative SE Bias} = \frac{MDN(\widehat{SE_r}) - MAD}{MAD}, \tag{8}$$

where MDN was the median of the estimated SE values and MAD was the empirical median-absolute-deviation of SE values. An absolute value of robust relative SE bias within 10% range was considered acceptable (**hooglandRobustnessStudiesCovariance1998?**).

Root Mean Squared Erorr (RMSE). The RMSE was defined as the squared root of the sum of squared bias:

$$RMSE = \sqrt{R^{-1}\Sigma_{r=1}^{R}(\hat{\gamma}_{xm_r} - \gamma_{xm})^2}. \tag{9}$$

RMSE measures the average difference between calculated interaction estimates and their true value, which can account for both bias, the systematic deviation from the true value, and variability, the spread of estimates across replications. In a 2,000 replication simulation, lower RMSE indicated greater accuracy in estimating $\hat{\gamma}_{xm}$. RMSE provided the most informative comparison across methodologies when key factors, including sample size, model complexity, and disturbance level, are held constant in the simulation.

Empirical Type I Error and Statistical Power. The empirical Type I error rate was computed as the proportion of replications in which the Wald test rejects H_0 at the significance level $\alpha = .05$ under the condition $\hat{\gamma}_{xm} = 0$. The empirical power was computed similarly for the condition $\gamma_{xm} = 0.3$.

Results