Comparing Methods for Adding Control Variables in Structural Equation Modeling

Suppanut Sriutaisuk & Sunthud Pornprasertmanit

Faculty of Psychology, Chulalongkorn University, Thailand, e-mail: suppanut.sri@gmail.com

BACKGROUND AND OBJECTIVE

Control variables or covariates usually refer to factors that are not of primary interest, but they are important to include in a model for several reasons. For example, in correlational studies, control variables are often added to provide more accurate estimates of the relationships between predictors and outcomes, or to rule out alternative explanations (Becker, 2005; Bollen & Bauldry, 2011). While issues on statistical control have been fairly established in traditional regression analyses, the same issues become convoluted when researchers adopt SEM in their studies. In the present study, we evaluated common methods using the Monte Carlo simulation.

METHOD

Data Generation and Analysis. In this study, outcomes are Rejection Rate (RR), Fit indices, parameter estimates and standard errors (*SE*). There are 6 conditions as follows:

1. Analysis models (Model): *Covariate → Latent, Covariate → Latent + Manifest, No Covariate*, and *Covariate → Manifest*
2. Factor loadings (Loadings): .40, .55, and .7
3. Covariate effects (Cov): .1/.2, .3/.4 and .5/.6 on the first and second factors respectively
4. Differential item functioning (DIF): 0, .1, .3, and .5 (on one manifest variable per factor)
5. Structural Coefficient (β): .1, .3, and .5
6. Sample Size (*N*): 100, 250, 500, and 1,000

We simulated and analyzed data from R (version 3.3.3; R Core Team, 2016) using the *simsem* (version 0.5-14; Pornprasertmanit, Miller & Schoemann, 2016) and *lavaan* (version 0.5-23.1097; Rosseel, 2012) packages. A total of 4 × 3 × 3 × 4 × 3 × 4 = 1,728 conditions were simulated and analyzed with 1,000 replications.

Population and Analysis Models. Figure 1 shows the population model structure (top-left), in which two latent variables (i.e., F_1, F_2), measured by four manifest variables (i.e., x_1–x_4 and y_1–y_4), are regressed on a covariate. Four analysis models were selected to reflect what we believed to be commonly used in practice: (a) a model without any covariates (*No Covariate*), (b) a model with direct paths from covariates to each factor (*Covariate → Latent*), (c) a model with direct paths from covariates to all factors and some manifest variables allowing measurement non-invariance or DIF (*Covariate → Latent + Manifest*), and (d) a model with direct paths from covariates to manifest variables but not latent variables (*Covariate → Manifest*).

Table 1 *Eta-Squared of the main and interaction effects from ANOVAs (1,605 conditions).*

	RR	RMSEA	Est Bias	SE Bias	Std Est Bias
Model	.288	.219	.561	.001	.534
Loadings	.008	.026	.015	.157	.001
Cov	.007	.015	.048	.011	.096
DIF	.228	.222	.019	.007	.007
β	.000	.000	.000	.019	.000
N	.028	.042	.011	.221	.000
Model:Cov	.011	.022	.141	.017	.135
Model:DIF	.263	.261	.134	.014	.107
Loadings:DIF	.012	.058	.002	.025	.000
Model:β	.001	.001	.002	.005	.035
Model:N	.035	.003	.000	.007	.001
Loadings:N	.001	.001	.007	.075	.000
Model:Loadings:DIF	.016	.075	.011	.024	.006
Model:DIF:N	.031	.004	.001	.008	.001

Note. $\eta^2 < .03$ are underlined. Rows in which all $\eta^2 < .03$ are omitted. However, all main effects are shown. RR = rejection rate; RMSEA = root mean square error of approximation; Est Bias = absolute difference of the unstandardized structural coefficient; SE Bias = relative difference of the standard errors of the unstandardized estimate; Std Est Bias = absolute difference of the standardized structural coefficient; Cov = Covariate effects; DIF = differential item functioning; β = the coefficient from the first to second factor.

Acknowledgement. Some parts of this project was developed under Prof. Moritz Heene at *Learning Sciences Research Methodologies* lab, Ludwig Maximilian University of Munich, Germany.

Population and Analysis Models

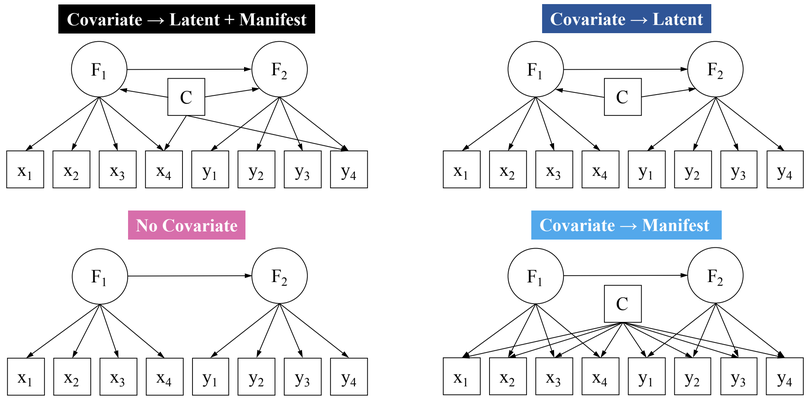

Figure 1. Path diagrams for population (top-left) and analysis models. F_i = factor or latent variable; C = covariate; x_i and y_i = indicator or manifest variable.

RESULTS

Seven percent of the conditions were excluded since they yielded a non-convergence result (4%) or returned less than 1% convergence rate (3%). The remained conditions had 933 converged replications on average (*SD* = 171, Median = 1,000). Table 1 shows selected main and interaction effects of each condition on outcomes.

Rejection Rate (RR) and RMSEA. The results are presented in Figure 2A and 2B. When measurement structures are invariant, the average rejection rate and RMSEAs of every analysis model is approximately at the desirable rate (.05 and near 0, respectively). When DIF ≠ 0, RR and RMSEA of the two misspecified models (i.e., *No Covariate* and *Covariate → Latent*) raise, while RR and RMSEA of other analysis models remained constant. Regarding the two misspecified models, sample size has noticeable effect on RRs, while RMSEAs are influenced by factor loadings. In conditions with which factor loadings are small, RMSEAs are below .06 even DIF is high.

Biases in structural unstandardized and standardized coefficients. We present graphical results in Figure 2C and 2D. When *Covariate → Latent* is specified and DIF is presented, the structural coefficient between latent variables is negatively biased. As expected, even DIF = 0, the omission of the covariate leads to the biased coefficient. The size of bias is mostly a function of the strength of the relationship between the covariate and latent variables. Although specifying *Covariate → Manifest* yields correct unstandardized coefficients across all conditions, standardized coefficients are problematic and misleading.

Standard error. Overall, *SE*s are not biased (relative bias < 10%) when sample size and/or factor loadings are high enough (e.g. *N* = 250 when factor loadings = .55). This pattern is similar across all analysis models.

DISCUSSION, LIMITATIONS AND RECOMMENDATIONS

The present study investigated methods for controlling control variables in SEM. Our recommendation is that researchers should regress factors on covariates, and may free more paths from covariates to manifest variables when necessary and theoretically support. Further, our results indicate that while the analysis model which included all possible direct paths from the covariate to manifest variables but not latent variables (*Covariate → Manifest*) yielded unbiased unstandardized estimates, the standardized estimates are misleading because the specification alters the meaning of the constructs. The construct is defined by unexplained observed variable variances rather than all observed variable variances. In social sciences, we frequently interpret parameter estimates in their completely standardized form. As such, specifying *Covariate → Manifest* variables or related methods (e.g., residual-centering) should be used with caution.

There are important limitations related to generalizability that should be noted. In the present study, every analysis model has only one main predictor and one outcome, which is unlikely in practice, where researchers normally adopt mediation and/or moderation hypotheses. Future research should study these types of model in SEM context. Also, in multilevel models where covariates can be controlled in one level or many levels.

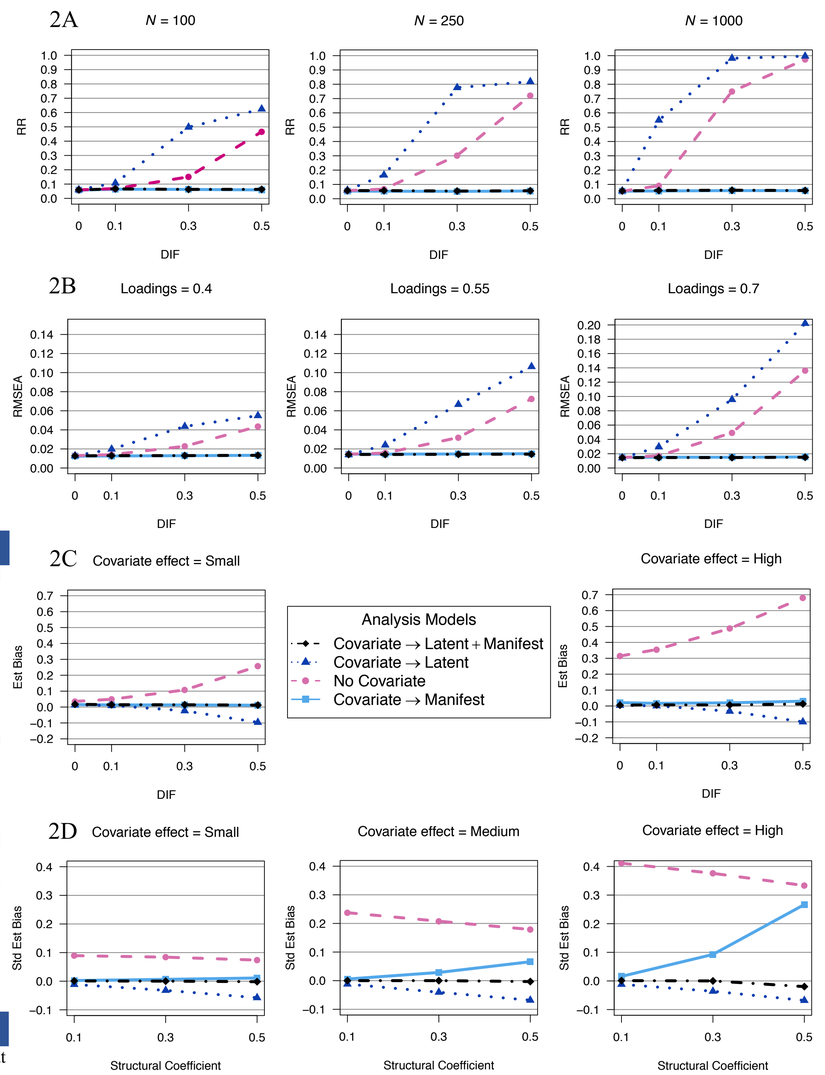

Figure 2. The average rejection rate (RR; Figure 2A), root mean square error of approximation (RMSEA; Figure 2B), absolute difference of the structural coefficient (Est Bias; Figure 2C-D) across different conditions; Std = standardized; DIF = differential item functioning; *N* = sample size.

REFERENCES

Becker, T. E. (2005). Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, *8*(3), 274-289.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychological methods*, *16*(3), 265-284.

Pornprasertmanit, S., Miller, P., Schoemann, A., Quick, C., & Jorgensen, T. (2016). Package simsem. Retrieved from http://cran.r-project.org/web/packages/simsem/simsem.pdf

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.