Supplementary Document for
"Summed versus Estimated Factor Scores: Considering Uncertainties When Using Observed Scores"

Yang Liu[1] Jolynn Pek [2]

Contents

[1] Department of Human Development and Quantitative Methodology, University of Maryland, College Park. Correspondence author. Email: yliu87@umd.edu

[2] Department of Psychology, the Ohio State University.

A Generation of Model Error

A.1 Cross Loadings

Model error due to misspecification was introduced by including additional cross loadings when generating data but omitting those cross loadings when fitting the model. The six cross loadings added to the baseline model (Figure 1) are: (a) from η_2 to y_2 and y_5, (b) from η_3 to y_8 and y_{11}, and (c) from η_1 to y_{14} and y_{17}. The six cross loadings are constrained to be equal within each condition. The common value of the cross loadings was selected to yield the desired population Root Mean Square Error of Approximation (RMSEA) ε, which is a function of the minimum value of the maximum likelihood (ML) discrepancy function F_{ML} (see Equation 18).

Let $\mathbf{\Sigma}_0(\kappa)$ be the $J \times J$ covariance matrix of the data generating model with cross loadings, in which $\kappa > 0$ denotes the common value of the six cross loadings; dependencies on other parameters are suppressed for notational succinctness. Also let $\mathbf{\Sigma}(\omega)$ be the $J \times J$ covariance matrix implied by the baseline model, in which ω denotes the q free parameters (i.e., factor loadings, inter-factor correlations, and unique variances).[3] The discrepancy between $\mathbf{\Sigma}_0(\kappa)$ and $\mathbf{\Sigma}(\omega)$ can be measured by

$$F_{\text{ML}}\left(\mathbf{\Sigma}_0(\kappa), \mathbf{\Sigma}(\omega)\right) = \log\det\left(\mathbf{\Sigma}(\omega)\right) - \log\det\left(\mathbf{\Sigma}_0(\kappa)\right) + \operatorname{tr}\left(\mathbf{\Sigma}_0(\kappa)\mathbf{\Sigma}(\omega)^{-1}\right) - J. \qquad \text{(S1)}$$

For each fixed $\kappa > 0$, let

$$F_0(\kappa) = \min_{\omega} F_{\text{ML}}(\mathbf{\Sigma}_0(\kappa), \mathbf{\Sigma}(\omega)) \qquad \text{(S2)}$$

be the minimized ML discrepancy function value with respect to ω. F_0 quantifies the distance of the fitted covariance structure to the generating one at the level of the population, known as the "discrepancy of approximation" (Cudeck & Henly, 1991). Given a target F_0 value representing a certain level of misfit, say $\tau > 0$,[4] the desired cross loading value can be found by solving

$$F_0(\kappa) = \tau. \qquad \text{(S3)}$$

[3] Common factor variances are fixed to 1 for model identification.

[4] F_0 is mapped to the population RMSEA ε via Equation 18.

When the model is sufficiently regular and κ and τ are sufficiently small, $F_0(\kappa)$ (Equation S2) is uniquely defined and the root finding problem (Equation S3) can be uniquely solved. In practice, a numerical search (e.g., Newton-type algorithms) can be employed to evaluate Equation S2 and solve Equation S3 approximately.

A.2 Random Noise

Cudeck and Browne (1992) described a method to randomly perturb a known population covariance structure such that the minimum discrepancy function value between the perturbed and original covariance matrices attains a prescribed value. Adapting our notation in Section A.1, Cudeck and Browne's (1992) approach aims to generate a random $\mathbf{\Sigma}_0$ such that

$$F_0(\mathbf{\Sigma}_0) = \min_{\omega} F_{\mathrm{ML}}(\mathbf{\Sigma}_0, \mathbf{\Sigma}(\omega)) = F_{\mathrm{ML}}(\mathbf{\Sigma}_0, \mathbf{\Sigma}(\omega_0)) = \tau. \tag{S4}$$

In Equation S4, $F_0(\mathbf{\Sigma}_0)$ denotes the minimized ML discrepancy function value, $\mathbf{\Sigma}(\omega)$ denotes the covariance structure implied by the baseline model, and τ is the prescribed minimum discrepancy function value as usual. In addition, let ω_0 be the known population parameter values under the baseline model. Equation S4 requires the minimized discrepancy between the randomly generated $\mathbf{\Sigma}_0$ and the family of covariance structures $\mathbf{\Sigma}(\omega)$ to be exactly τ; moreover, the minimum is attained at the known population parameter values ω_0.

To achieve Equation S4, it is noted that the gradient of F_{ML} with respect to ω is supposed to vanish at the minimum $\omega = \omega_0$ when ω_0 is in the interior of the parameter space and F_{ML} is sufficiently smooth. Cudeck and Browne (1992) showed that the gradient can be expressed by

$$\left.\frac{\partial}{\partial \omega} F_{\mathrm{ML}}(\mathbf{\Sigma}_0, \mathbf{\Sigma}(\omega))\right|_{\theta=\omega_0} = \mathbf{B}(\omega_0)' \mathrm{vecs}\left(\mathbf{\Sigma}_0 - \mathbf{\Sigma}(\omega_0)\right), \tag{S5}$$

in which vecs$(\cdot)$ extracts the non-duplicated entries in a symmetric matrix, and the $p^* \times q$ matrix $\mathbf{B}$ is a function of ω where $p^* = p(p+1)/2$. It follows that $\omega = \omega_0$ is a stationary point of $F_{\mathrm{ML}}(\mathbf{\Sigma}_0, \mathbf{\Sigma}(\omega))$ if

$$\mathbf{\Sigma}_0 - \mathbf{\Sigma}(\omega_0) = \gamma \mathbf{E}, \tag{S6}$$

in which $\gamma > 0$, and $\mathbf{E}$ is a $p \times p$ symmetric matrix such that the $p^* \times 1$ vector $\mathbf{e} = \text{vecs}(\mathbf{E})$ lies in the null space of $\mathbf{B}(\omega)$; i.e., $\mathbf{B}(\omega_0)'\mathbf{e} = \mathbf{0}$. A stronger result was also established by Cudeck and Browne (1992) in that the stationary point ω_0 is in fact a global minimum when γ is not too large.

By Equation S6, simulating $\mathbf{\Sigma}_0$ that satisfies Equation S4 can be achieved in two steps:

1) generate a random vector $\mathbf{e}$ residing in the null space of $\mathbf{B}(\omega_0)$ and obtain the corresponding $\mathbf{E}$;

2) find the value of γ (numerically) such that $F_{\text{ML}}(\mathbf{\Sigma}(\omega_0) + \gamma\mathbf{E}, \mathbf{\Sigma}(\omega_0)) = \tau$.

There are various ways to generate random directions from the null space of $\mathbf{B}(\omega_0)$. In our simulation, the random vector $\mathbf{e}$ is determined as follows:

1) Generate a $J \times J$ matrix $\mathbf{A} \sim \text{Wishart}_J(\mathbf{I}_J, J)$ where $\mathbf{I}_J$ is a $J \times J$ identity matrix;

2) Regress $\text{vecs}(\mathbf{A})$ on $\mathbf{B}(\omega_0)$ and let $\mathbf{e}$ be the residuals of the regression. Stated differently, set $\mathbf{e} = \left[\mathbf{I}_{p^*} - \mathbf{B}(\omega_0)\left(\mathbf{B}(\omega_0)'\mathbf{B}(\omega_0)\right)^{-1}\mathbf{B}(\omega_0)'\right] \text{vecs}(\mathbf{A})$.

The sum of squared residual covariances (*SSRC*) statistic computed in Table 2 was the sum of squared off-diagonal entries in $\gamma\mathbf{E}$, which measures the degree of model misspecification. Because our $\mathbf{\Sigma}(\theta)$ is determined by a common factor model, it follows that

$$\mathbf{\Sigma}_0 = \mathbf{\Lambda\Phi\Lambda}' + \underbrace{\mathbf{\Theta} + \gamma\mathbf{E}}_{\mathbf{\Theta}^*}. \tag{S7}$$

It is tempting to conclude based on Equation S7 that Cudeck and Browne's method introduces model error as unaccounted dependencies among unique factors, quantified by the non-diagonal matrix $\mathbf{\Theta}^* = \mathbf{\Theta} + \gamma\mathbf{E}$. While this claim is conceptually true, note that $\mathbf{\Theta}^*$ is not guaranteed to be positive definite; hence, we are not always able to generate unique factors with $\mathbf{\Theta}^*$ being the covariance matrix. In our simulation, we performed accept-reject sampling and only keep $\mathbf{\Theta}^*$'s that are positive definite. Non-positive definite instances happened rarely and the vast majority of generated $\mathbf{\Theta}^*$ were retained.

B Regression Factor Scores

Recall that the classical test theory (CTT) true scores for regression factor scores (Equation 8) are expressed in Equation 11, which is reproduced below:

$$\mathbb{E}(\underline{\hat{\eta}}_i^R|\eta_i) = \mathbf{M}\eta_i = \underbrace{(\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{\Lambda} + \mathbf{\Phi}^{-1})^{-1}\mathbf{\Lambda}'\mathbf{\Theta}^{-1}\mathbf{\Lambda}}_{\mathbf{M}}\eta_i. \tag{S8}$$

Equation S8 suggests that the CTT true scores for regression factor scores are computed by pre-multiplying the matrix $\mathbf{M}$, which depends on the factor model parameters, to the latent variables (LVs) η_i. Under the baseline model (Figure 1) in our simulation study, the matrix $\mathbf{M}$ under the low, high, and wide communality conditions are

$$\mathbf{M}_{\text{low}} = \begin{bmatrix} .715 & .054 & .016 \\ .054 & .623 & .184 \\ .016 & .184 & .633 \end{bmatrix}, \; \mathbf{M}_{\text{high}} = \begin{bmatrix} .935 & .017 & .001 \\ .017 & .889 & .069 \\ .001 & .069 & .894 \end{bmatrix},$$
$$\mathbf{M}_{\text{wide}} = \begin{bmatrix} .891 & .027 & .003 \\ .027 & .826 & .104 \\ .003 & .104 & .833 \end{bmatrix}, \tag{S9}$$

respectively. All three matrices in Equation S9 have non-zero off-diagonal entries; therefore, each component of $\mathbb{E}(\underline{\hat{\eta}_i}^R|\underline{\eta}_i)$ is not perfectly correlated with the corresponding component of $\underline{\eta}_i$. To see this, we first compute the covariances between $\mathbb{E}(\underline{\hat{\eta}_i}^R|\underline{\eta}_i)$ and $\underline{\eta}_i$:

$$\text{Cov}\left(\mathbb{E}(\underline{\hat{\eta}_i}^R|\underline{\eta}_i), \underline{\eta}_i\right) = \text{Cov}(\mathbf{M}\underline{\eta}_i, \underline{\eta}_i) = \mathbf{M}\mathbf{\Phi}. \tag{S10}$$

Then we arrive at the correlation matrix $\text{Corr}(\mathbb{E}(\underline{\hat{\eta}_i}^R|\underline{\eta}_i), \underline{\eta}_i)$ upon normalizing rows and columns of Equation S10 by the corresponding standard deviations of $\mathbb{E}(\underline{\hat{\eta}_i}^R|\underline{\eta}_i)$ and $\underline{\eta}_i$, respectively. We denote the resulting correlation matrices by $\mathbf{P}_{\text{low}}$, $\mathbf{P}_{\text{high}}$, and $\mathbf{P}_{\text{wide}}$ under the three communality conditions and present their numerical values as follows:

$$\mathbf{P}_{\text{low}} = \begin{bmatrix} .9963 & .3797 & .2768 \\ .3582 & .9835 & .8086 \\ .2627 & .8132 & .9846 \end{bmatrix}, \ \mathbf{P}_{\text{high}} = \begin{bmatrix} .9998 & .3178 & .2232 \\ .3164 & .9985 & .7363 \\ .2223 & .7367 & .9986 \end{bmatrix},$$

$$\mathbf{P}_{\text{wide}} = \begin{bmatrix} .9995 & .3298 & .2328 \\ .3261 & .9962 & .7558 \\ .2304 & .7567 & .9966 \end{bmatrix}. \tag{S11}$$

Even though correlations between CTT true scores and the corresponding LVs are not perfect (i.e., diagonal elements of the three matrices in Equation S11) for regression factor scores, they are in fact very close to one. Therefore, it remains approximately true that a CTT reliability analysis reveals how regression factor scores mirror true LVs under the baseline independent cluster model (Figure 1) in our simulation experiment.

C Additional Results for the Industrialization and Democracy Data

Reliability coefficients and the associated 90% bootstrap confidence intervals (CIs) are tabulated in Table S1. The pattern remains similar to the Holzinger-Swineford example. Both Bartlett and regression scores tend to have slightly higher reliabilities compared to the summed score. Here, the coefficient alpha is no longer a lower bound for the coefficient omega for the two democracy index factors: This is a known fact due to non-zero covariances among unique factors.

Table S1: Reliabilities (90% bootstrapped confidence intervals [CI] based on 2000 bootstrapped samples) for the three-factors in the Industrialization and Democracy example. Note that alpha for summed scores was computed based on the sample covariance matrix, whereas omega for summed scores and reliabilities for Bartlett and regression scores were computed from Equation 13. ind60: Industrialization latent variable (LV) at 1960. demo60: Democracy LV at 1960. demo65: Democracy LV at 1965.

Factor	alpha	90% CI	omega	90% CI	Bartlett	90% CI	regression	90% CI
ind60	0.9	[.88, .92]	0.94	[.92, .96]	0.97	[.95, .99]	0.97	[.95, .99]
demo60	0.86	[.81, .90]	0.84	[.77, .89]	0.88	[.84, .92]	0.92	[.89, .95]
demo65	0.88	[.84, .91]	0.86	[.80, .90]	0.89	[.84, .92]	0.93	[.91, .96]

References

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika*, *57*, 357–369. doi: 10.1007/BF02295424

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519. doi: 10.1037/0033-2909.109.3.512