Reconciling factor-based and composite-based approaches to structural equation modeling

Edward E. Rigdon (erigdon@gsu.edu)
Modern Modeling Methods Conference
May 20, 2015
Thesis:

Arguments for factor-based SEM being “better than” composite-based approaches (particularly PLS path modeling) are invalid.
Factor-based SEM

- Common factors A, B, C
- Observed variables \( a_1 \ldots a_j, b_1 \ldots b_k, c_1 \ldots c_m \)
- Factor / factor: factor / indicator

\[
C = g_1 A + g_2 B + e
\]

\[
a_i = \lambda_{iA} A + e_{ia}, i = 1, j
\]
\[
b_i = \lambda_{iB} B + e_{ib}, i = 1, k
\]
\[
c_i = \lambda_{iC} C + e_{ic}, i = 1, m
\]
PLS path modeling

- Composites A, B, C
- Observed variables $a_1 ... a_j$, $b_1 ... b_k$, $c_1 ... c_m$
- “Inner model”
  \[ C = g_1 A + g_2 B + e \]

- “outer model”
  \[ A = \sum_j w_{ai} a_i \]
  \[ B = \sum_k w_{bi} b_i \]
  \[ C = \sum_m w_{ci} c_i \]
- With routine standardization
Mode A, Mode B

• “Mode B”: regression-weighted sum of indicators
• “Mode A”: correlation-weighted sum of indicators, ignoring collinearity among predictors (Dana & Dawes 2004, Waller & Jones 2010)
• Regardless of specifics, PLS (like GSCA) uses weighted composites where factor-based SEM uses common factors
“The tidy theory of error laid down for psychology at the start of the century by Spearman and Brown has always seemed just a little too tidy to describe the perverse behavior of real data.”

Cronbach, Gleser, Nanda & Rajaratnam (1972), p. v
Measurement Model

Conceptual Variable

Proxy

(Un)reliability

(In)validity

Mathematical Operations

Observed Variables
Thesis:

Arguments for factor-based SEM being “better than” composite-based approaches (particularly PLS path modeling) are invalid.
The arguments

• Bias: Factor-based SEM yields unbiased parameter estimates, while PLS parameter estimates are biased.

• Measurement error: Factor-based SEM accounts for measurement error, while PLS does not.

• Latent variables: Factor-based SEM involves “latent variables,” while PLS does not.

• Overall fit: Factor-based SEM assesses overall fit, while PLS does not.
1. Factor-based SEM yields unbiased parameter estimates, while PLS estimates are biased

• Wold (1982): PLS employs an “intentional approximation”


• The nature of Dijkstra and Henseler’s (2015) “consistent PLS” implies that bias exists.
Counter: Bias is due to model misspecification, not intrinsic in PLS path modeling

• Rigdon (2012) speculated that, for a correctly specified population, PLS parameter estimates would be at least consistent.

• Becker et al. (2013) specified a composite-based population, and found PLS parameter estimates to be consistent, converging on specified values as n increases.

• BUT for small n, bias can be substantial

• (PLS also has its own “identification” requirements)
Population model
Becker et al. (2013, in prep)
\[
\Sigma = \begin{bmatrix}
\Sigma_{YY} & \Sigma_{XY}' \\
\Sigma_{XY} & \Sigma_{XX}
\end{bmatrix}
\]
Specifying a composite-based population
Becker et al. (2013, in prep)

• $\Sigma_{YY}$, $\Sigma_{XX}$: Specify “within” covariances for each set of components
• $W, V$: Specify weights, which define composite variances
• $P$: Specify “structural” path weights (and / or covariances) between composites
• $\Sigma_{XY}$: Calculate “across” covariances consistent with desired path weights, using path tracing rules $\Sigma_{XY} = \Sigma_{YY} V'PW\Sigma_{XX}$
Sample results

Figure 3. Path Coefficient Estimation Error for Different Sample Sizes
2. Factor-based SEM accounts for measurement error, while PLS does not

- Factor-based SEM separates common factor from “specific factor” or residual, which has been called “measurement error” (Jöreskog 1983)
- Bollen (1989) explicitly associated “common factor” with “true score,” thus linking residuals with measurement error
- PLS composites share in any “error” variance contained in their components, to some extent
Counter: common factors retain an “error component” via factor indeterminacy

• De-meaned factor model with \( p \) indicators \( Y \), \( k \) common factors \( F \):
  \[
  Y = \Lambda \ F + I \ Z
  \]

• Covariance model, common factors orthogonal to specific factors:
  \[
  \Sigma = \Lambda \ \Phi \ \Lambda' + I \ \Theta \ I
  \]

• Consolidating \( \Phi, \Theta \) into \( \Psi \):
  \[
  \Sigma = \begin{bmatrix} \Lambda & I \end{bmatrix} \begin{bmatrix} \Phi & 0 \\ 0 & \Theta \end{bmatrix} \begin{bmatrix} \Lambda & I \end{bmatrix}' = \begin{bmatrix} \Lambda & I \end{bmatrix} \Psi \begin{bmatrix} \Lambda & I \end{bmatrix}'
  \]

• Rank \( \Sigma \) = \( p \). Rank \( \Psi \) = \( p + k \).
  (Guttman 1955)
The indeterminate error component in common factors

\[ F = \Phi \Lambda ' \Sigma^{-1} Y + PS \]

\[ PP' = M = \Phi - \Phi \Lambda ' \Sigma^{-1} \Lambda \Phi \]

Schönemann and Steiger (1976)
Factor indeterminacy and “error”

• Guttman’s (1955) determinacy metric: \[ \rho_{\text{min}} = 2 \times \rho_{F,Y}^2 - 1 \]
  where \( \rho_{F,Y}^2 \) is R\(^2\) for this common factor predicted by all observed variables in the model.

• Mulaik 2010, p. 380: \( \rho_{F,Y}^2 \) calculated from model parameters:
  \[ P_{F,Y}^2 = \text{diag} \left( \Phi \Lambda \Sigma^{-1} \Lambda \Phi \right) \]

• McDonald (1999, p. 89): for a single common factor, \( \rho_{F,Y}^2 \) equals McDonald’s \( \omega \), i.e., “composite reliability” — BUT ONLY when loadings are equal.
Values of Guttman’s $\rho_{\text{min}}$

<table>
<thead>
<tr>
<th>Loading</th>
<th>Number of Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>.5</td>
<td>-.200</td>
</tr>
<tr>
<td>.6</td>
<td>.059</td>
</tr>
<tr>
<td>.7</td>
<td>.315</td>
</tr>
<tr>
<td>.8</td>
<td>.561</td>
</tr>
<tr>
<td>.9</td>
<td>.790</td>
</tr>
</tbody>
</table>
Factor-based SEM does not eliminate error—it repackages error as indeterminacy.

Neither approach to SEM truly eliminates error, except at the limit.
Indeterminacy does not affect estimation of relations within the factor model

Indeterminacy does blur relations with variables outside the model

(Steiger 1979)
The conceptual variable is outside the model.
3. Factor-based SEM involves latent variables, while PLS does not

• “Latent variable” can mean either:
  (statistical) common factor, or
  (substantive) conceptual variable

• Obviously, factor-based SEM involves factors, while composite-based SEM does not
Counter: conceptual variables are important to both

• Either approach can be used to learn about conceptual variables . . .
• . . . though not all applications of either method are well-suited to that purpose
• Without conceptual variables and validated proxies, multiple indicator methods may be just fancified data description
4. Factor-based SEM assesses overall fit, while PLS does not

• No doubt, factor-based SEM has an overall $\chi^2$ statistic, and composite-based PLS path modeling does not

• (Composite-based generalized structured component analysis (GSCA) claims an overall $R^2$ criterion)

• But does it matter?
Measurement Model

Conceptual Variable

Proxy

(Math)validity
(Un)reliability

Does $\chi^2$ fit tell us anything about the quality of the proxy thus created?

Mathematical Operations

Observed Variables
“If a value of $\chi^2$ is obtained, which is large compared to the number of degrees of freedom, this is an indication that more information can be extracted from the data.”
Jöreskog (1969, p. 201)

In itself, the existence of more information in the data says nothing about the validity of a given proxy (though indeterminancy will mean unreliability)
P_F might be a great proxy for conceptual variable A but not so much for B. \( \chi^2 \) fit of the statistical model will be the same, either way.
Overall fit assessment in factor-based SEM does not make factor proxies “better” than composite proxies.
None of these arguments offers a valid basis for always preferring a factor proxy over a composite proxy.
So, what would make one kind of proxy better than another, in a given situation?
What does all this mean?

• Learning about conceptual variables can be done with either factor-based SEM or composite-based SEM, or possibly a combination (via Dijkstra and Henseler’s “consistent PLS”)
• BUT we must enable meaningful assessment of our proxies
• Besides using large and relevant samples and properly accounting for data distributions
• We must take full advantage of prior information, and embrace the discipline of the scientific imagination (Feynman, 1994)
We also need to think about the value of “exploratory” multiple indicator analyses, where nothing is known about the conceptual variables supposedly involved.
“If we take the simile of the bridge crossing a river by way of an island, there is a statistical span from the near bank to the island, and a subject-matter span from the island to the far bank. Both are important.”

Cornfield & Tukey (1956), p. 913
Thank You!

Questions
References