**The Role of Visualizations in SEM**

**Evaluation of SEMs**

The typical approach to evaluating SEMs is quite different than those typically used in generalized linear modeling. Whereas there is greater emphasis in SEM than GLM in visual model specification through path models or directed acyclic graphs (DAGs), there is less emphasis on visualization during the model evaluation stage. GLM textbooks emphasize residual diagnostic procedures to evaluate the tenability of model assumptions (e.g., normally distributed errors, homoskedasticity, etc.) and evaluation of model predictive performance using plots of predicted vs. observed values in addition to the overall F statistic, and R2 metrics (Fox, 2008). Reporting of global fit indices is ubiquitous in SEM applications, but visual model inspection (post-estimation) seems to be rare. This is likely in part attributable to important differences in GLM and SEM modeling objectives.

GLM is used to develop both causal and predictive models. In the predictive modeling context, parsimonious GLMs that account for much of the variance in outcomes are desirable. Residuals reflect the difference in observed (*Y*) and model-predicted () values of the outcome variable. SEM, in contrast, is typically a causal modeling approach. The goal is to test theory by specifying a model that accurately represents real-world causal processes. Residuals reflect discrepancies between model-implied and sample covariances and are indicative of a misspecified model. Although some scholars have emphasized the importance of attending to residual covariances in model evaluation (Goodboy & Kline, 2017), there appears to be far greater focus on global evaluation strategies. In both GLM and SEM frameworks, the ability to obtain an accurate estimate of the causal effect of an exposure on an outcome (i.e., a causal estimand) is predicated on the specification of a model that neutralizes confounding influences.

A SEM encodes a researcher’s beliefs about the causal processes underlying the observed data (i.e., the data-generating mechanisms). Naturally, evaluating whether there is empirical evidence against one’s hypothesized model is critical prior to accepting the model’s worldly implications as reasonable. If the hypothesized causal model provides a strong representation of the real-world data-generating mechanisms—and the study procedures and statistical assumptions are sound—then we expect the model-implied covariance matrix to closely resemble the sample covariance matrix. If the discrepancy is in excess of what we might reasonably expect from sampling variability alone, there is evidence against the hypothesized model.

There is an extensive literature—and impassioned debate—over the best ways to evaluate the adequacy of SEMs. The traditional method of evaluating the global fit in overidentified models is to conduct a χ2 test to determine if there is evidence of discrepancy between the model-implied covariance matrix () and the sample covariance matrix (**S**). A significant χ2 test indicates a model-data discrepancy that is exceeds chance expectation (Bollen, 1989). A common criticism of the χ2 test is that it is sensitive to sample size. , with large sample sizes, even very small discrepancies between and **S** will lead to model rejections. As researchers generally view models as imperfect approximations of reality, many argue that strict adherence to the χ2 test will result in the elimination of imperfect but potentially useful models that depart from the true model in trivial ways. Consequently, SEM scholars have developed dozens of approximate fit indices (AFIs) evaluating the degree of discrepancy or the performance of the hypothesized model compared relative to a null model. Simulation studies have led to the adoption of conventional cutoff values for AFIs above (for goodness of fit indices) or below (for badness of fit indices) which is indicative of acceptable correspondence between the model and sample data (Hu & Bentler, 1999).

Unfortunately, the magnitude of model-data discrepancies is not necessarily an indicator of how consequential model misspecifications are. Models that provide a “close fit” to the data—and even perfectly fitting models —can badly misrepresent the data generating mechanisms (L. A. Hayduk, 2014). There are no fixed cutoff values for AFIs that can reliably differentiate meaningful and trivial model misspecifications (Chen et al., 2008). Consequently, some scholars have called into question the value of the AFIs, urging SEM practitioners to regard all evidence of misspecifications that exceed conventional tolerance for chance expectation as being potentially indicative of model deficiency (Barrett, 2007; L. A. Hayduk, 2014; McIntosh, 2007).

**Previous Approaches to Visualizing LVMs**

Stem-and-leaf plots are a common visual aid for evaluating local fit (Bollen, 1989). Plotting the distribution of the residual covariances (or correlations) in a stem-and-leaf plot can aid in giving an overall impression of the magnitude of the model-data discrepancies and the identification of outlying elements with particularly large residuals. These plots may help modelers focus in on specific elements of the model that are least consistent with the observed data. Stem-and-leaf plots are particularly helpful for complex models in which the residual covariance matrix has many elements, making visual inspection tedious.

Muthén and Asporouhov (2017) demonstrated the utility of visual diagnostic methods—reminiscent of those used in linear regression models—in detecting certain types of model misspecifications. First, they showed that plots of estimated factor scores for a latent outcome variable against observed predictor variables can be used to detect unspecified nonlinear effects of the predictor on the latent outcome. Second, they used residual scatterplots to detect violations of local independence in a latent factor model. When two reflective indicators (Y1 and Y2) of a latent factor (η) had an unmodeled common cause, they showed that plotting the indictors’ residuals against one another could help identify the non-zero residual covariance, prompting modelers to consider an alternate model. Finally, they demonstrated in a latent factor model how plotting predicted values for a reflective indicator () against the observed indicator values (*Y*) could uncover unmodeled heterogeneity that could be better captured using a mixture model. A benefit of these visual diagnostic strategies relative to commonly used tests of model fit and approximate fit indices is that the plots may provide guidance in determining the nature of the model misspecifications and how they might be remediated (Asparouhov & Muthén, 2017).

There are notable limitations, however, to the Asporouhov and Muthén (2017) approach. First, when using factor score estimates for model diagnostic purposes, the result may differ depending on the method of factor score estimation owing to factor indeterminacy (Grice, 2001; Rigdon et al., 2019; Steiger, 1996). At high levels of indeterminacy, two sets of factor score estimates could differ markedly—and even be negatively correlated (Grice, 2001). Thus, estimating the degree of factor indeterminacy would be prudent prior to relying on the factor score estimates and potentially evaluating whether similar visual patterns hold across multiple methods of calculating factor score estimates. Second, as the authors note in their article, in their method of obtaining predicted values for an outcome variable (), the latent factor scores () contributes to the prediction even though already contains information about *Y*. However, this dependence between *Y* and becomes less problematic when the latent factor is measured with increasing numbers of highly reliable indicators (Asparouhov & Muthén, 2017).

* Briefly mention Wang et al (2005) residual diagnostic for GMMs
* problem that they have to reorient (Bollen, 1989)