Published in final edited form as:
Soc Sci Res. 2022 September ; 107: 102769. doi:10.1016/j.ssresearch.2022.102769.

FIFTY YEARS OF STRUCTURAL EQUATION MODELING: A HISTORY OF GENERALIZATION, UNIFICATION, AND DIFFUSION

Kenneth A. Bollen[1,2,3,4], **Zachary Fisher**[5], **Adam Lilly**[1,2,4], **Christopher Brehm**[1,2,4], **Lan Luo**[3,4], **Alejandro Martinez**[3,4], **Ai Ye**[3,4]

[1]Carolina Population Center

[2]Department of Sociology

[3]Department of Psychology and Neuroscience

[4]University of North Carolina, at Chapel Hill

[5]Pennsylvania State University

INTRODUCTION

In 1972, the birth year of *Social Science Research* (*SSR*), Structural Equation Modeling (SEM) was experiencing a decade-long rebirth, moving from its origins in genetics to the social and behavioral sciences. University of North Carolina (Chapel Hill) sociologist H.M. Blalock (1964 [1961]) published his highly influential book on *Causal Inferences in Nonexperimental Research*. Sociologist Otis Dudley Duncan (1966) reinforced and extended the awareness of Sewall Wright's (1921) path analysis technique, demonstrating how path analysis could further the understanding of causal relations between variables. Soon SEM spread even further in the social and behavioral sciences. Blalock's (1971) edited book *Causal Models in the Social Sciences* collected causal inference papers from a variety of disciplines. Goldberger and Duncan's (1973) *Structural Equation Models in the Social Sciences* grew out of a 1970s conference and furthered this multidisciplinary examination of SEM and its use in the social and behavioral sciences. As Henry Riecken wrote in the forward to the 1973 book:

> *The Conference on Structural Equation Models is a milestone in the history of social science methodology, for it brought together the widest variety of social scientists whose research interests lay in the development and use of quantitative methods for analyzing causation in non-experimental data. The conferees came from economics, education, political science, psychology, and sociology. Their topical interests ranged across occupational mobility, educational achievement, Congressional voting, conflict resolution, and macroeconomic policy."* Goldberger and Duncan, (1973, p. xiii)

This rebirth of SEM was marked by three striking trends. The first two trends were the generalization and unification of common statistical methods into a single model. The SEM publications from the 1960s and 1970s blurred the lines between econometrics, psychometrics, and sociometrics. They showed how a general model could incorporate simultaneous equation models from econometrics and sociology while also controlling for measurement error and multiple indicators. It placed factor analysis as a special case of the same framework. Concomitantly, it provided structural or causal interpretation of parameters when analysts input causal assumptions. The third trend was the diffusion of SEM through numerous fields. This facilitated the realization that many traditional models fell under a more general model which in turn opened the possibility of new hybrid models.

The generalization and unification that was part of SEM and the diffusion of SEM across scientific fields took off in the 1960s and 1970s and has continued through today. Our paper has several purposes: (1) to provide a historic sketch of SEM with an emphasis on developments over the last 50 years, (2) to illustrate contemporary applications of SEM using recent papers in *SSR*, (3) to present some strengths and vulnerabilities of SEM, and (4) to exemplify areas ripe for further development.

The next section gives a historical sketch of SEMs, including their origins. Following this section, we present the general nature of SEM and its assumptions. After this are sections on contemporary uses and developments in SEM and on its strengths and vulnerabilities. Subsequent sections are on frontier topics, followed by a conclusion.

HISTORICAL SKETCH OF SEMs

If generalization, unification, and diffusion are trends that mark the last half-century of SEMs, this was less true at its origins. If we focus on the measurement model component of contemporary SEM, Spearman's (1904) factor analysis is a natural starting point. He sought to test whether a general intelligence factor underlies performance on a variety of tests. If we think about path analysis, Sewall Wright (1918; 1921) is its originator. His earliest efforts were anchored in biology and genetics. Wright's (1918) first study examined bone size in rabbits. Indeed, the bulk of Wright's applications were to biological and genetic examples. However, some of his publications illustrated the potential applications of path analysis to economics (Wright, 1925; Wright, 1928[1]) and psychology (Wright, 1934). But the widespread diffusion of SEM that occurred during the last 50 years departed from the isolated examples of the first 50 years.

The last 50 years are also marked by a misunderstanding that equates path analysis to models with only observed variables and error terms. To discover the fallacy in this thinking, one need only read Wright's earlier publications. Wright (1918) examined the size components of rabbit bones. Using partial correlations, he suggested a general size factor as well as factors specific to each group of bones. In contemporary terms, the general size and group-specific factors are latent variables, and the model he described corresponded to a bifactor model (see Bollen, 1989a: pages 5–6). Wright (1921) noted the applicability of path analysis even with "hypothetical factors" (page 575). A few years later, Wright (1925) applied path analysis to time series data to study corn and hog correlations. This

[1] *The Tariff on Animal and Vegetable Oils* was authored solely by Philip G. Wright, Sewall Wright's father, but letters between the two of them uncovered by James Stock and Kerry Clark show that the ideas in Appendix B were jointly developed by both Wrights (Angrist & Pischke, 2015).

time series model gave a prominent role to hog breeding as a latent variable (or hypothetical factor). Wright (1934, pages 165–167) gave a path analysis justification for the correction for attenuation when measurement error and latent variables were present. Hence, latent variables in path analysis were present at its origin and were not an innovation of the last 50 years.

Indeed, other key components of contemporary SEMs were visible early on though many aspects have been generalized and formalized. To briefly expand on these points, we use the steps widely associated with contemporary SEM: 1) model specification, 2) model implied moments, 3) identification of parameters, 4) estimation, and 5) model fit. We sketch what has been long present and how the last 50 years have differed.

Model Specification

Consider model specification via path diagrams first. One of Wright's fundamental contributions was to specify models through graphs that contained variables and single- and double-headed arrows which incorporated causal and noncausal assumptions. In the last 50 years, these diagrams have evolved with widespread, but not universal, meanings assigned to the symbols. Latent variables are typically enclosed in an ellipse or circle, observed variables are in boxes, and the disturbances (errors) are often unenclosed in either. The single-headed arrow signifies a direct causal effect from the variable at the base of the arrow to that at the head of the arrow (e.g., $X_1 \rightarrow Y_2$ means X_1 has a direct causal effect on Y_2). The double-headed arrow symbolizes an unanalyzed association between the two variables it links (e.g., $X_1 \leftrightarrow X_2$ states that X_1 and X_2 are correlated for unspecified reasons). The single-headed arrows typically correspond to equations that are linear in the parameters. This is not to say that nonlinear relationships are absent in classic SEM. For instance, Winship and Mare (1983) and Muthén (1984) presented path diagram notation for models with categorical endogenous variables that are nonlinear in the linkages of the observed variables to their underlying counterparts. Bollen (1995) proposed path diagram notation for models that were nonlinear in the variables, but linear in the parameters. Claims that classic SEMs consists only of "linear" relationships are decades out of date even if these claims persist (e.g., VanderWeele, 2012).

Wright's proposal to represent causal relations in diagrams also gave birth to Directed Acyclic Graphs (DAGs) that have grown in popularity over the last couple of decades (e.g., Pearl, 2000). DAGs use some of the same symbols that Wright proposed but sometimes have different meanings. In classic SEM, the single-headed arrow (e.g., $X_1 \rightarrow Y_2$) means a linear in parameter relationship between the two variables, even if the variables enter an equation nonlinearly (e.g., interaction or squared term). In DAGs a specific functional form (or distribution) is unspecified, variables are called nodes, and the single-headed arrow is called a directed edge between two nodes. DAGs are sometimes referred to as Nonparametric SEMs or NPSEM (Pearl, 2000) because of this general form. Furthermore, the DAG literature altered the meaning of the two-headed arrow (e.g., $X_1 \leftrightarrow X_2$), from an unanalyzed association to an association caused by an omitted variable (Pearl, 2000, pg. 12). Other than latent confounders, the DAG literature has largely focused on observed rather than latent variables. Elwert and Winship (2014) present an example DAG that includes a latent variable and its imperfect measure, but this is rare. This contrasts with classic SEM that routinely includes latent variables, measurement error, and multiple indicators. When viewing causal diagrams, readers need to be sure they know the intended meaning, and whether the author is using classic SEM path diagrams or DAG notation. However, these two descendants of Wright have in common the use of graphs to make explicit the causal assumptions of the researcher and to easily communicate which variables influence others and which do not.

Beyond the developments in more general path diagram notation, the last 50 years gave rise to more general model notation for equations. The LISREL notation of Jöreskog (1973) is well-known and closely related to that of Keesling (1973) and Wiley (1973). There is a matrix equation that specifies the relationship between the latent variables and two matrix equations for the measurement model. We write the latent variable equation as,

$$\boldsymbol{\eta} = \boldsymbol{\alpha}_\eta + \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

where $\boldsymbol{\eta}$ is a vector of the latent endogenous variables, $\boldsymbol{\alpha}_\eta$ is a vector of equation intercepts, $\boldsymbol{B}$ is the matrix of regression coefficients giving the effects of each latent endogenous variable on the others, $\boldsymbol{\Gamma}$ is the coefficient matrix for the latent exogenous variables in $\boldsymbol{\xi}$, and $\boldsymbol{\varsigma}$ is the vector of errors or disturbances. The model assumes $\mathrm{E}(\boldsymbol{\varsigma}) = \mathbf{0}$ and $\mathrm{C}(\boldsymbol{\varsigma}, \boldsymbol{\xi}^{'}) = \mathbf{0}$. The measurement model is,

$$\boldsymbol{y} = \boldsymbol{\alpha}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{x} = \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}$$

where $\boldsymbol{y}$ is a vector of indicators of the latent endogenous variables $(\boldsymbol{\eta})$, $\boldsymbol{\alpha}_y$ is an intercept vector, $\boldsymbol{\Lambda}_y$ is a matrix of factor loadings, and $\boldsymbol{\varepsilon}$ is the vector of errors for the indicators. The model assumes $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{C}(\boldsymbol{\varepsilon}, \boldsymbol{\eta}^{'}) = \mathbf{0}$. The vector $\boldsymbol{x}$ contains the indicators of the latent exogenous variables $(\boldsymbol{\xi})$ with analogous definitions and assumptions of the remaining terms.

This and similar notation enabled SEM to incorporate many common statistical models as special cases. For instance, if we assume no measurement error ($\boldsymbol{\alpha}_y$, $\boldsymbol{\alpha}_x$, $\boldsymbol{\varepsilon}$, $\boldsymbol{\delta}$ are all zero and $\boldsymbol{\Lambda}_y$, $\boldsymbol{\Lambda}_x$ are identity matrices), this model is equivalent to simultaneous equation models that are well-studied in econometrics (Jöreskog, 1973). Furthermore, if there is only a single dependent variable, then this is equivalent to multiple regression and if the covariates consist only of dummy variables, then the model specializes to ANOVA. If instead we only use the equation $\boldsymbol{x} = \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}$, we can represent all factor analysis models with or without correlated errors (unique factors). Growth curve models, random and fixed effects longitudinal models, autoregressive and cross-lagged panel models, and multitrait-multimethod models are just a few of the other models that this notation can represent.

This general notation and model made connections between seemingly unrelated methods and disciplines using more specialized formulations. It helped to bring together econometric, psychometric, and sociometric models (see e.g., Goldberger & Duncan, 1973). It provided a framework that could include Spearman-like factor analyses and the earlier path analysis models of Wright and his successors with or without latent variables, multiple indicators, and measurement error and opened the door to new model structures.

Model Implied Moments

Sewall Wright employed path diagrams along with his tracing rule to decompose the correlation between variables into the parameters of the specified model. He demonstrated in many specific examples how the correlations were dependent on the parameters of the model. His approach was to specify a model and then to derive the parameters that made up each correlation of the observed variables. Blalock (1961, 1963, 1964), Duncan (1966, 1975), and other sociologists who helped to spread path analysis into the social and behavioral sciences took a similar approach: specify a particular model and decompose each correlation, variance, or covariance into the parameters of that model.

The general Jöreskog-Keesling-Wiley model that we described in the last section took a matrix approach. Rather than freshly deriving the correlations, variances, and covariances of the observed variables for each new example, they proposed the general model to derive these implied moments in matrices. By doing so, they decomposed the moments of the observed variables into the model parameters for any model that fell under this general notation. To illustrate this, consider the measurement equation $\boldsymbol{x} = \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}$ and its model implied covariance matrix of

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x^{'} + \boldsymbol{\Theta}_\delta$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the model implied covariance matrix of the indicators in $\boldsymbol{x}$, $\boldsymbol{\theta}$ contains all the factor loadings and variances and covariances of the factors ($\boldsymbol{\xi}$) and errors that are free parameters in the model, $\boldsymbol{\Phi}$ is the covariance matrix of the latent factors, and $\boldsymbol{\Theta}_\delta$ is the covariance matrix of the errors (unique factors). For any specific factor analysis model, users can substitute the factor loading matrix, covariance matrix of the factors, and covariance matrix of the errors into the right-hand side of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Doing so provides an expression for all variances and covariances in terms of the model parameters. This is also true for a model that includes the latent variable and measurement models using the expanded notation for the general model. This ability to generate the implied moments for any model is another reflection of the generalization and unification of SEMs that occurred during the last 50 years.

Identification of parameters

Identification refers to whether there exists a unique set of parameter values that satisfy the implied moment equations for a specific model. Typically, the means, variances, and covariances of the observed variables are the only observed variable moments that analysts use. In Wright's early applications he focused on correlations ("standardized covariances"). He wrote each correlation of the observed variables as a function of the parameters in the specific model under consideration. Then he demonstrated that each model parameter could be uniquely solved in terms of the correlations or variances of the observed variables. If there was more than one way to use these correlations and variances to solve for the same parameter, the parameter was overidentified. If there was no unique solution available, then the parameter was not identified (underidentified).

Researchers still apply this algebraic approach to the identification of some models, but within the last 50 years more general rules of identification have been proposed. For instance, the Two Indicator Rule (Kenny, 1979; Bollen, 1989a) is a sufficient condition for identification for factor analysis and it requires that each latent variable has at least two indicators, each factor correlates with at least one other factor, no indicator loads on more than one factor, and there are no correlated errors. The Recursive Rule for simultaneous equations (no measurement error) is a sufficient condition for identification and it assumes the errors of all endogenous variables are uncorrelated and there are no feedback relations in the model. Similarly, there are other rules of identification in the literature (e.g., Bollen, 1989a). These cover many but not all models and hence generalize and unify the approach to identification compared to the early path analysis applications that treated each specific example uniquely. Determining whether the information matrix of maximum likelihood estimation is nonsingular is a common empirical method to check model identification. However, this is a test of local identification which is necessary but not sufficient for global identification (Bollen, 1989, pages 246–251). Hence establishing global identification is preferable when feasible.

Estimation

The previous section described early efforts to identify models. These efforts involved showing a parameter could be uniquely solved as a function of the correlations or variances of the observed variables. If there was more than one such function for the parameter, then it was overidentified. Wright (1921, 1934) and those who followed would often use the equations that established identification in terms of population correlations (or covariances, variances) of the observed variables for estimation, replacing the population quantities with their sample counterparts. When there was more than one solution for the same parameter, the estimates would be combined by averaging. The last 50 years were marked by more sophisticated estimation that applies to general SEMs such as the LISREL model described above. Just as the general model permitted formulation of a broad range of models with and without latent variables, these general estimation procedures would not require new estimation methods for each new, specific application. Land (1973), for instance, demonstrated the desirable properties of a maximum likelihood (ML) estimator for fully recursive models of observed variables and its superiority to ad hoc ways of combining estimates of overidentified recursive models. Jöreskog and Sörbom’s (1974) LISREL program was a turning point in that it made an ML estimator widely available for SEM. This generalization and unification of estimation meant an incredibly broad range of models had a sophisticated estimator available which did not need to be recreated for each individual model. Furthermore, estimation capabilities for SEM have expanded to handle categorical and count endogenous variables, mixture models, multilevel models, hazard rate equations, and more. Furthermore, a wide variety of estimators for SEM including

Bayesian and instrumental variable estimators are available. Not only did the general SEM model specification incorporate and unify many statistical models, but a general estimation framework also greatly increased the popularity of SEM. SEM software has capitalized on these developments, rapidly spreading from the early beginnings in LISREL to widespread availability in major general statistical software programs such as Mplus, Stata, SAS, SPSS, and R.

Model Fit

As previously stated, in many of Wright's applications he derived the relationship between the parameters of the model and the correlations of the observed variables. He used these to estimate parameters. With overidentified parameters, he could compare the different estimates of the same parameter from the different solutions. The closer these different estimates of the same parameter were, the greater was the consistency between the specified model and the data. These comparisons were an informal gauge of model fit.

In the last 50 years, a much wider variety of fit measures have emerged, spreading through publications and their inclusion in SEM software. The likelihood ratio chi square test statistic that compares the model implied moments (e.g., implied covariance matrix) to the sample moments (e.g., sample covariance matrix) is the best-known fit statistic. In its usual form it tests the null hypothesis of $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$ where $\mathbf{\Sigma}$ is the population covariance matrix of the observed variables, $\mathbf{\Sigma}(\boldsymbol{\theta})$ is the population model implied covariance matrix, and $\boldsymbol{\theta}$ is a vector containing the parameters of a model. If the model is valid, then the equality should hold. A statistically significant chi square suggests at least one part of this equality fails, indicating the model has a specification error.

The chi square is a *statistical* significance test of Ho: $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$ and it is not a test of the substantive magnitude of a misspecification. In large samples, the statistical power of the test can be large and even substantively minor omissions are detectable. Many fit statistics to supplement the chi square test have emerged since the 1970s. Some are based on comparing the chi square and degrees of freedom of the hypothesized model to that of a very restrictive baseline model (e.g., Bentler, 1990; Bollen, 1989b; Tucker & Lewis, 1973). Others are standalone indexes that rely only on the model chi square and its degrees of freedom, relating these to other quantities such as a noncentral chi square distribution (e.g., Steiger and Lind, 1980; Steiger 2016) or approximations to Bayes Factors (Raftery, 1995). The guidelines for and interpretation of fit statistics are subject to controversy, but common practice is to report and assess both the chi square test statistic as well as several fit indices.

In addition to these fit measures that assess the overidentification constraints and model fit, there are more familiar ways to assess components of fit such as the magnitude and statistical significance of the parameter estimates. When the model implied instrumental variable estimator is employed, there also is an equation specific test for each overidentified equation (Bollen, 1996, 2019; Kirby & Bollen, 2009).

Summary

At its origins, SEM was applied example-by-example where the parameters were related to the correlations of the observed variables based on the context of a particular application. Identification and estimation were *ad hoc* and based on solving each parameter as a function of the correlations, covariances, or variances of the observed variables. When there was more than one solution, these estimates were combined. Besides the specific application orientation, biological and genetic examples dominated with little expansion to other substantive areas.

The last 50 years have departed from these origins. Path diagrams have evolved. A general model notation that incorporates the most widely used statistical models is widespread. Models that are linear and nonlinear in the variables are part of classic SEM and DAGs represent a general nonparametric form of SEM. General rules of identification have developed that cover numerous possible structures rather than a single application. Similarly, general estimation and model fit methods have been created and implemented in widely available software. All these features have led to the diffusion of SEM far beyond its origins in biology and genetics or the measurement of intelligence in psychology. In brief, the last 50 years have been marked by generalization, unification, and diffusion.

In the next section, we illustrate contemporary uses as reflected in *SSR*.

CONTEMPORARY USES AND DEVELOPMENTS

As we described in the last section, SEM uses a general model that incorporates many specific models. Applications of SEM make use of different techniques subsumed by the general model to address diverse research questions. To provide a sense of this diversity, we conducted a search for the term “structural equation model” using the database of articles published in *SSR*. We reviewed the abstracts of the *SSR* articles that were published from 2011–2022 and selected SEM articles that demonstrated the breadth of SEM. While less comprehensive than a true systematic review, this simple search strategy yielded many interesting and varied applications of SEM in *SSR*.

We begin with the measurement model and examples of how researchers used SEM to measure abstract concepts, to explore measurement invariance, to validate new measures, and to predict membership in latent classes. We then discuss applications of the latent variable or structural model to investigate multiple outcomes, mediation, and the evaluation of causal hypotheses. We end the section by illustrating applications of SEM to longitudinal data, to reciprocal effects, and in combination with inverse probability of treatment weighting (IPTW).

The first study focuses on measuring abstract concepts. Researchers use confirmatory factor analysis (CFA), a special case of SEM, to analyze the relations between latent variables that represent these concepts and their measures or indicators (see, e.g., Roos & Bauldry, 2022). In one study, Beierlein et al. (2016) sought to measure “group-focused enmity”, which they define as hatred, animosity, or prejudice towards minority groups. The prejudices they measure include sexism, anti-Semitism, xenophobia, homophobia, Islamaphobia, and anti-homelessness. They first estimated a measurement model for each dimension (sexism, anti-Semitism, etc.) where the latent prejudices were hypothesized to influence the indicators and explain the association among them. Then two second-order factors, anti-immigrant as well as sexual prejudice, were introduced to explain the first-order factors of xenophobia and Islamaphobia, and sexism and homophobia, respectively. Finally, a third-order factor labeled group-focused enmity was introduced to explain the two second-order factors and the first order factors of anti-homelessness and anti-Semitism. Their analysis illustrates how SEM can test models with latent variables at distinct levels of abstractness as well as those latent variables that directly influence the indicators.

When we compare the average values of a scale in two or more groups, we implicitly assume the indicators of the scale relate to their respective latent variables in the same way. For example, we might ask whether men or women experience more depressive symptoms by comparing average scores on the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1977). This comparison requires that the scale measures depression in an equivalent way for both men and women.[2] If the relations between measures and their latent variables differ across groups and researchers ignore this, then regression coefficients and even mean values are biased and misleading. Multiple group SEM allows researchers to examine invariance and test measurement invariance by formally comparing the measurement model parameters between two or more groups. Raudenská (2020) investigates the measurement invariance of a set of categorical indicators of positive and negative affect across 22 European countries, and two time points. The analysis begins with the standard multiple group SEM approach of testing exact measurement invariance but scalar invariance is rejected. Partial scalar invariance is achieved by relaxing equality constraints on thresholds sequentially, guided by modification indices, but this approach may not lead to the correct partial invariance model due to the substantial number of possible modifications (Kim et al., 2017). Raudenská (2020) addresses this by using a Bayesian approach to test for approximate invariance which involves setting a small prior variance for the between-group parameter differences (Muthén & Asparouhov, 2013). After rejecting approximate scalar invariance based on the Bayesian approach, the estimated differences between each threshold and its average across the countries are used to identify a set of invariant items for each of the two constructs.

Seddig and Lomazzi (2019) view the lack of measurement invariance as something to be explained. Their focus is the lack of invariance when measuring gender role attitudes across countries. The authors use multilevel CFA, where the indicator variance is decomposed into within- and between-country variance. Because their gender role indicators lack scalar invariance, they focus on the error variance in the between-level equation, which must not be significantly different from zero for scalar invariance to hold in multilevel models. To explain the noninvariance, they introduce the cultural value embeddedness and gender inequality indices as country-level predictors into the equations for the group-specific item intercept and the between-level latent factor. They find that the lack of invariance is primarily explained by the cultural value embeddedness index, suggesting that

[2]It also assumes that depression is a one dimensional variable, a questionable assumption that we ignore here (e.g., Perreira et al., 2005).

cultural differences between countries influence responses to items measuring gender roles independently from the constructs they were designed to measure.

Another use of SEM is to evaluate new indicators and scales. Meuleman et al. (2020) designed the Control, Attitude, Reciprocity, Identity, and Need (CARIN) scale to measure citizens' deservingness opinions, a concept which had been discussed in theoretical literature but only measured with proxies. They reason that citizens evaluate the deservingness of potential welfare recipients before forming opinions about policy, and their deservingness is evaluated according to the five principles included in the CARIN scale name. They use SEM to investigate the dimensionality, validity, and reliability of the scale, and test its construct validity by whether it predicts policy preferences. This illustrates a role for SEM in both assessing old measures and developing new ones.

Our discussion so far has focused on *continuous latent* variables, but sometimes the latent variable is *categorical*. Latent class analysis (LCA) is useful in situations where we want to use information from the observed indicators to predict membership in unobserved classes or categories.[3] Branic and Hipp (2018) use LCA to classify neighborhoods in Southern California into latent classes based on the changes in the socioeconomic characteristics of the neighborhoods and their residents during the 10-year period between 2000 and 2010. They first created two-year change scores using home mortgage loan data on home purchase, refinance, and improvement loans. The variables collected for each loan type were the average loan amount, the number of loans granted, and the loan type's percentage of all granted loans. They additionally collected the number of borrowers moving into the neighborhood each year by income level. To reduce the number of variables, Branic and Hipp (2018) conducted principal component analysis to combine the indicators into five composite measures. These five composite change scores measured at eleven time points entered the LCA as observed variables. A model with eleven classes provided the best solution, where for example the "urban investors" class included neighborhoods with higher income borrowers moving in and with a lagging increase in home improvement loans. Based on these and other characteristics, the authors characterize these neighborhoods as ones that were under revitalization or gentrification. The concept of a gentrifying neighborhood is familiar to most people, but this study demonstrated that multiple factors contribute to whether a neighborhood is gentrifying. This example illustrates how LCA can classify observations into categories that while not directly observable, are meaningful.

The previous examples illustrated *SSR* SEM applications that focus on the relation between latent variables and their indicators. However, SEM also allows models where latent variables directly and indirectly influence other latent variables. Mediation analysis was central to early SEMs and understanding these and similar mechanisms continues to motivate current SEM applications. Using SEM to assess mediation is popular because not only can researchers obtain point estimates and standard errors for indirect effects, but model fit statistics are available to assess the fit of overidentified models. Furthermore, latent variables and missing data are easily handled. Wallace (2012) uses SEM mediation techniques to test whether fear of crime mediates the relationship between perceptions of

[3]See Bauer and Curran (2004) for a cautionary note on the use of LCA.

neighborhood disorder and three outcomes: self-rated health, depression, and anxiety. The author considers fear of being robbed/attacked, fear of a home invasion, and fear of walking alone at night. While there are other available techniques for testing mediation, the benefits listed above, as well as the ease of including multiple mediators as is done in Wallace (2012), motivates the continuing use of SEM for this purpose.

In addition to testing for measurement invariance, multiple group SEM allows testing whether parameters from the latent variable ("structural") model are the same across groups. Researchers can consider unequal coefficients due to interactions, as well as allow variance and covariance parameters to differ across groups. Pudrovska et al. (2014) use this approach in their examination of life-course mechanisms linking early-life socioeconomic status (SES) to body mass index (BMI) in later life. Their models specify direct effects from early-life SES to BMI at age 65, as well as indirect effects operating through health behaviors and socioeconomic resources at various points in the life course. They then use multiple group analysis to test whether these path coefficients differ significantly between men and women.

The ability to assess a hypothesized model's fit to the data is a feature of SEM that is common to the general model and all special cases of it, as we mentioned earlier. The model chi square test and common fit indices all reflect the correspondence of the model implied and the observed covariance matrices. In an overidentified model, there will be multiple ways to solve for one or more parameters using the information in the covariance matrix and therefore the model implied and observed covariance matrices will not be equal as they are in exactly identified models. If there is a larger discrepancy than would be expected due to sampling error, this is evidence that the model is incorrect. For this reason, many contemporary users of SEM select among competing models based on their fit. For example, Bollen, Gutin, Halpern, and Harris (2021) considered three different measurement models to explain the dimensionality and reliability of four indicators of self-rated health, including two contemporaneous measures, a retrospective measure, and a parental report measure. Their initial model includes no equality constraints on the factor loadings, while the second model constrains the factor loadings of the contemporaneous measures to be equal. A third model allows the measurement errors of the retrospective measure and parental report to covary, a modification suggested by the Sargan test statistic, an equation-level fit statistic which accompanies the MIIV-2SLS estimator (Bollen, et al., 2021). The authors use the global chi-square test, BIC, CFI, TLI, and RMSEA to compare the three models, and all measures of fit indicated the third model as the best fit.

SEM also permits a multitude of longitudinal models. One example is the autoregressive cross-lagged panel model, which primarily consists of two variables measured at multiple time points. To ground the discussion, we refer to Yucel and Borgmann's (2022) study of the reciprocal relationship between work-to-family conflict (WTFC) and depressive symptoms. Each of these variables is measured at three time points and the autoregressive part of the model refers to the effect of depressive symptoms at one time point on depressive symptoms at the subsequent time point, and likewise for WTFC. Yucel and Borgmann (2022) also include a direct effect of each variable at time one on the same variable at time three. The cross-lagged part of the model refers to the fact that WTFC influences depressive symptoms at the subsequent wave and depressive symptoms also influence WTFC at the subsequent wave. This model exploits the longitudinal structure of the data and allows for the estimation of the reciprocal effects of WTFC and depressive symptoms on one another while controlling for the lagged dependent variable in each equation.

Another SEM application is to nonrecursive models, or models with feedback or reciprocal effects. Boggess's (2017) *SSR* paper investigates the reciprocal relationship between changes in a neighborhood crime rate and changes in the proportions of racial/ethnic groups living in the neighborhood. Changes in the violent and property crime rates are instrumented by changes in income inequality and in the percent of workers who are service workers. The author argues that changes in income inequality will not directly influence decisions to move out of the neighborhood because changes in the level of inequality are not obvious to neighborhood residents, but that income inequality has been shown to influence crime. Similarly, the change in the percentage of service workers is expected to influence crime because workers in the secondary market have less to lose when choosing to commit a crime, and the service industry connects potential offenders and targets. To instrument racial/ethnic change, Boggess (2017) uses change in the percentage of crowded households and change in the average home value. The author argues that the change in the proportion of crowded households will necessarily change the racial/ethnic composition, but that it is unlikely to influence crime through a different mechanism. Changes in property values are also expected to directly influence both in- and out-migration from a neighborhood. The ability to estimate reciprocal effects with cross-sectional data is a powerful component of SEM, but it relies heavily on the researcher's ability to select instrumental variables that meet the assumptions.

Continuing its historical trajectory of the incorporation of techniques from other fields into SEM analyses, recent SEM applications have begun to incorporate techniques from the counterfactual tradition of causal inference. For example, Shen et al. (2021) incorporate IPTW into their SEM analysis on the influence of father's absence on the educational attainment of children in China. They use SEM to estimate a series of indirect effects operating through economic capital, human capital, social capital, cultural capital, behavioral problems, and educational achievement. IPTW is employed to deal with selection into father absence, which is a binary indicator. IPTW creates a pseudo-population where the confounder distributions should be similar between the treatment and control groups and has benefits over matching where observations that cannot be matched must be omitted. IPTW and other counterfactual methods are also designed to overcome problems related to misspecification of the parametric form of the influence of the confounders on the other variables in the model.

In the past ten years of *SSR*, there were many other applications of SEM that we did not cover. Beyond *SSR*, there are SEM analyses regularly published in other journals both within and outside of the social sciences. The generality of the model has facilitated its diffusion across many substantive fields and allowed for the generalization and unification of more specific methodological techniques. There are many other SEM techniques in use that were not demonstrated by the examples above. Latent growth curve models are popular for analyzing longitudinal data that complement the autoregressive models discussed above, and there are further extensions that combine the two (Bianconcini & Bollen, 2018; Bollen & Curran, 2006). SEMs are no longer limited to continuous variables and the well-known maximum likelihood estimator. Estimators for categorical endogenous variables (Muthen, 1984) are widely available, and Bayesian SEM (Kaplan & Depaoli, 2012) and model-implied instrumental variable estimators (Bollen et al., 2021) provide both new approaches to estimation and new ways of understanding SEM.

STRENGTHS AND VULNERABILITIES

The last section on applications of SEM in *SSR* illustrated features available in SEM. In this section we explicitly highlight some strengths and vulnerabilities of SEM, of which researchers should be aware.

Strengths

Making Models Explicit Rather Than Implicit—Too often the results of empirical articles have only tables with a dependent variable and covariates, their coefficients, and standard errors. These coefficients are the direct effects estimates of the independent variables on the outcome variable. What we do not know from these tables is the nature of the relations among the covariates with each other or the presence of latent variables or correlated errors, where unstated assumptions can create confusion about the nature of the model itself. An advantage of SEM is that the path diagram and the equations make model specification explicit. It forces the researcher to ask which variables are exogenous and which are endogenous and how they relate to each other; which variables have minimal measurement error; which are latent variables that require multiple indicators to successfully measure; when are equation or measurement errors uncorrelated or correlated. The beauty of path diagrams is that these assumptions become explicit, allowing readers to readily assess their plausibility. In contrast, a table without a model leaves much unknown.

Mediation Analysis (Direct, Indirect, and Total Effects)—Classic SEM pioneered mediation analysis by drawing distinctions between direct, indirect, and total effects. A highlight of Wright's (1934) path analysis was that it provided these decompositions explicitly. In the last 50 years, mediation has received considerable attention with a focus on estimation and testing (e.g., Alwin & Hauser, 1975; Baron & Kenny, 1986; Sobel, 1982), including extensions to latent variable SEM (Jöreskog & Sörbom, 1981) and the ability to estimate and test specific effects through any subset of variables in the model (Bollen, 1987). In addition, the SEM framework has been extended to handle nonlinear and nonparametric mediation analysis (Pearl, 2012; Muthén, 2011; Bollen & Pearl, 2013), allowing for a wider range of functional forms.

Measurement of Latent Variables—Concepts or constructs in the social and behavioral sciences are often abstract. Whether it is social capital, depression, or socioeconomic status, it is rarely possible to measure the latent variables that represent these ideas without measurement error. SEM enables researchers to build models that relate the latent variables to multiple observed indicators. We gave examples of this in the section on current applications. Using latent variables and their observed indicators we can test the dimensionality of measures. We can estimate how well these indicators measure the latent variable and obtain estimates of their reliability and validity. We can test whether the errors of two or more indicators correlate. Furthermore, as illustrated in the application section, it is possible to test measurement invariance across different groups. In other words, SEM encourages the specification of measurement models to account for the fact that many constructs in the health, social, and behavioral sciences are not directly observable.

Taking Account of Measurement Error—Closely related to the measurement of latent variables is the ability of SEM to estimate the relationships between latent variables while simultaneously controlling for measurement error. No matter what the variable is, whether as concrete as currency or temperature, or as abstract as behavior or mood, the measurement often contains both random and nonrandom errors. When not accounted for, these errors can bias regression estimates or estimates of other parameters (Bollen, 1989a; Rigdon 1994). SEM recognizes this imperfect nature of measurement by explicitly specifying measurement error, while conventional multiple or categorical regression models, multilevel analyses, or related methods implicitly assume that the explanatory variables (covariates) have negligible measurement error. Rather than simply ignoring this measurement error, SEM allows us to account for it while simultaneously testing whether various assumptions about these measurements hold.

Tests of Causal Assumptions—Researchers do not derive causal relations from SEMs; a SEM represents and relies upon the causal assumptions input by the researcher (Wright, 1921; Bollen & Pearl, 2013).[4] The strongest of these assumptions are represented by omitted paths in path diagrams and missing edges in NPSEMs (e.g., DAGs), and zero associations between errors and/or exogenous variables which implies no omitted confounding variables. In essence, we assume that the graph (path diagram or DAG) is an accurate representation of the causal relations. When a SEM has overidentified parameters, we can test empirical implications that should hold if the model is valid. Some of these, such as the likelihood ratio test that we introduced earlier, test the overidentification constraints in the *whole* model simultaneously. It is a global test. The overidentification tests of equations that accompany MIIV-2SLS are *local* tests of the overidentified equations (Bollen, et al., 2021), as are tests of partial correlations (Blalock, 1964) and tests of local independence in DAGs. Vanishing tetrad tests (Blalock, 1964; Bollen & Ting, 1993) can test *parts of* or the *whole* model. All these tests provide evidence of the consistency between the causal assumptions and the data, but they cannot prove the validity of the causal assumptions. In other words, like with common null hypothesiss tests, we can reject the causal assumptions but cannot fully validate them.

The specification of the model and the empirical tests of whether these specifications are consistent with the data are logically prior to and separate from the estimation of the magnitude of the causal effects. Classic SEM has a variety of estimators for this latter task that depend on the nature of the model and whether the variables are continuous or not.[5]

[4]In the section on Future Directions, we describe exploratory methods aimed at causal discovery. Though even here, some causal assumptions are necessary.

Other Advantages—The generality of SEM provides additional advantages that we touch on briefly. One is the number of diagnostics that have been developed. If we can incorporate a more specialized model into SEM, we can take advantage of these diagnostics when they might not be otherwise available. For instance, Meredith and Tisak (1990) demonstrated that latent growth curve models are a special type of CFA in SEM. With this, it was possible to estimate these models with SEM software and to generate fit statistics and tests that were not available with traditional methods of estimation. Similarly, traditional and new forms of random and fixed effects longitudinal models are accessible within SEM, providing diagnostics and test statistics not otherwise available (Allison & Bollen, 1997; Bollen & Brand, 2010). The use of these statistics is not limited to models with continuous endogenous variables, as SEM has expanded to incorporate binary or ordinal endogenous variables (Muthén, 1984; Jöreskog & Sörbom, 1985).

Vulnerabilities

While SEM possesses many strengths, like all modeling techniques, it also has vulnerabilities. In this section, we briefly note several vulnerabilities. Specifically, we have chosen to organize these vulnerabilities into three categories: weak substantive theory, issues with model testing, and philosophical ambiguities. These categories do not constitute a comprehensive list of vulnerabilities but are important to highlight.

Model Specification and Weak Theory/Substantive Knowledge—Typically, a SEM analysis begins with the specification of a model. Specification depends on knowledge or speculations about the key latent variables and appropriate measures of them. Furthermore, the model should consist of theoretically derived hypotheses about which direct and indirect relations do and do not exist. Unfortunately, theory or substantive knowledge are rarely detailed enough to completely inform the specification of a model (Eronen & Bringmann, 2021; Fried, 2020; Klein, 2014; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019). Decisions regarding the dimensionality of constructs, functional form of relationships, the inclusion of correlated errors, and the omission of causal paths can easily become overwhelming. In practice, researchers tend to rely on their field’s modeling traditions to compensate for gaps in theory or substantive knowledge or apply statistical diagnostics and post-hoc justifications, which have their own challenges and shortcomings when used to support model respecification (Cliff, 1983; Silvia & MacCallum, 1988; Hoyle & Panter, 1995).

One of these modeling traditions is the tendency to assume a factor analysis structure for a set of observed variables. A strength of SEM is its ability to accommodate measurement errors. The strong tradition in test theory (Lord & Novick, 2008) and factor analysis (Spearman, 1904) is to assume that latent variables drive changes in their indicators. However, alternative structures should be considered. These could come in several forms, such as casual indicators where observed variables act as causes of a latent variable (Blalock, 1964; Bollen & Lennox, 1991; Bollen & Diamantopoulos, 2017), or network

[5]Estimation in the DAG tradition differs from SEM in that in classic SEM an explicit equation with coefficients is the norm even if the equations include nonlinear relations. DAGs are nonparametric so that an explicit equation and parameters are not part of the specification unless additional parametric assumptions are later introduced.

models where observed variables are a part of a complex network of mutually interacting relationships (Borsboom & Cramer, 2013). Even at the latent level one should consider that observed variables may be related to an unobserved categorical variable as in latent class analysis (McCutcheon, 1987).

Theory will never provide comprehensive guidance for model specification but the integration of machine learning with SEM, described below in the Future Directions section, offers a more principled approach to data-driven model specification.

Issues With Model Testing—The SEM framework offers researchers a flexible way to test theories, where models found to be inconsistent with the data are deemed to be inconsistent with reality (Bollen, 1989a p. 67–72). While on the surface this process seems straightforward, in practice it is more complicated. For one, there remains a debate in the literature concerning the role the chi square test and alternative assessments of fit should play when evaluating a model. Also, these measures of fit are sometimes influenced by model size, sample size, variable distributions, and other characteristics unrelated to the validity of the structure of the model. Finally, consistency of the model with the data is necessary but not sufficient for the consistency of the model with reality. Indeed, there may exist plausible alternative models that fit the data equally well (a phenomenon we go on to describe below) or fit better.

As noted previously, researchers can statistically evaluate a model's inconsistency with the data using the chi square test where the null hypothesis represents equivalence between the population observed variable covariance matrix and the model implied covariance matrix. However, the role that the chi square test should play in SEM continues to be debated amongst researchers. The special issue on SEM in *Personality and Individual Differences* (Vernon & Eysenck, 2007) showcases several perspectives on this debate. Some argue that rejection of the null hypothesis should result in immediate rejection of the model and lead to investigation of the source of misfit (Hayduk et al., 2007; Hayduk, 2014). Because the chi square test is powered to detect minor misspecifications when the sample size is large, others argue that rejection of the null hypothesis should be expected and tolerated because models are approximations to reality (Barrett, 2007; MacCallum, 1990). It was this perspective that engendered the development and use of many alternative methods for evaluating model fit deemed alternative fit indices (AFIs).

AFIs serve as largely descriptive tools for evaluating model fit, where interpretation of their values is aided by proposed cutoffs to serve as *guidelines* (Hu & Bentler, 1999). While never intended to become gold standard thresholds of performance, they have unfortunately become entrenched in modeling practice (Marsh et al. 2004). A further complication with the use of AFI cutoffs is that their performance differs based on research conditions, such as sample size (Chen et al., 2008), the number of variables in the model (Kenny & McCoach, 2003), magnitude of the correlations amongst the observed variables (Fornell & Larcker, 1981; Marsh et al. 2004), and measurement quality (McNeish, An, & Hancock, 2018; Browne et al., 2002). When working from the cutoffs alone, conflicting messages can create ambiguity associated with model evaluation. To address these problems, McNeish and Wolf (2021) provide a simulation-based method for determining application specific AFI cutoffs in CFA which considers characteristics of the model and data and is available in a convenient web-based Shiny application.

Another issue that further complicates the process of model testing is the phenomenon of chi square equivalent models. These are models with different specifications that provide identical model fit (Bollen, 1989a, pages 68–71; MacCallum, Wegener, Uchino, Fabrigar, 1993; Raykov & Marcoulides, 2001; van Bork et al., 2021). Several approaches have been developed to identify such models (e.g., Hershberger, 1994; Lee & Hershberger, 1990; Stelzl, 1986; Raykov & Penev, 1999). Thus, researchers who have specified a model that exhibits adequate fit must also consider alternative models which fit equally well but may differ dramatically in their theoretical implications. For these models, researchers must rely on criteria such as temporal precedence, research design, data collection, or theoretical plausibility to eliminate competing equivalent models. Of course, given the number of alternative models, and the limits to these criteria, it can become difficult or impossible to rule out the possibility of one of the alternatives being the true model. However as noted by Maccallum et al. (1993), the existence of equivalent models can be beneficial in its own right; guiding researchers towards new theoretical models to explore or rule out.

Philosophical Controversy in SEM—All methodological approaches are subject to philosophical debate and controversy and SEM is no exception. Two features we highlight are the nature of causality and latent variables. The nature of causality has long been a topic of discussion by both philosophers and scientists, but a unified definition still eludes us (Cook, Campbell, & Day, 1979; Pearl & Mackenzie, 2018). Examples of controversies surrounding causality are whether experimentation or manipulation is necessary for causation (Bollen, 1989a, Chapter 3; Holland, 1986; Morgan & Winship, 2015), whether the reduction of complex social phenomena to mathematical equations can capture causal processes (Cartwright, 1999; Rogosa, 1987; Wright, 1920), and what are the best methods for investigating causal phenomena (Rosenbaum and Rubin, 1983; Heckman, 2005). SEM was not developed to discover nor prove causation, but instead to serve as a tool for combining qualitative causal assumptions and empirical data to yield quantitative causal conclusions and measures of fit to assess the plausibility of the assumptions (Bollen & Pearl, 2013). Researchers therefore must navigate their research while considering these open inquiries into the nature of causation.

Another common feature associated with SEM is the inclusion of latent variables. This topic too contains its own ambiguities that need navigating. Latent variables have long been a part of SEM, with Wright's (1918, 1921) earliest work including them. However, how best to conceptualize and define latent variables is still open to discussion as several formal and non-formal definitions exist (Bollen, 2002; DeYoung & Krueger, 2020; Nunnally, 1978; Lord, 1953). The most inclusive view of latent variables is the *sample realization* definition, where one can define latent variables as researcher-specified variables with no sample realization (Bollen, 2002). Others reserve the term "latent variable" to indicate a statistical construct, represented by the shared portion of variance from a set of indicators (DeYoung & Krueger, 2020). Some philosophical discussions on latent variables have debated whether conceptualizing latent variables outside of their mathematical representation in SEM requires a realist interpretation of latent variables (e.g., Borsboom, Mellenbergh,

& van Heerden, 2003). In addition, several methodological issues, such as assessing measurement invariance, differential item and test function, local independence, and effect heterogeneity, to name a few, intersect with these philosophical positions. For this reason, many philosophical issues have practical implications for applied modeling. This is but a small subset of the views that have emerged from the discussion on latent variables. Without consensus on how to define and conceptualize them, applied researchers using SEM are left to operate in this ambiguity.

Summary of Strengths and Vulnerabilities—In this section, we highlighted several strengths and vulnerabilities of SEM. Most of the strengths are ones that are not found elsewhere, other than special cases of SEM. In contrast, many of the weaknesses are ones shared by most approaches. For instance, the alternative to latent variables is to assume that the relations hold among observed variables. This implicitly assumes no or negligible measurement error in our observed variables, a highly implausible assumption. Defining and establishing causality has been controversial for centuries no matter what technique is used. Finally, limited theory and substantive knowledge hinders all approaches. Thus, the vulnerabilities of SEM are widely shared.

FUTURE DIRECTIONS

As the last 50 years of SEM demonstrate, it remains an area of innovation and expansion. We cannot know SEM’s future, but here we feature a couple of areas that hold promise and are related to machine learning-computational applications and time series analysis. These are hardly the only areas of SEM development, but they are noteworthy in both their continuity with past research and their potential for taking SEM in new directions.

Machine Learning and Computational Methods in SEM

With the era of big data and artificial intelligence, machine learning methods have become popular in virtually all scientific fields. As a general statistical modeling method, SEM is no exception to this trend. One of the vulnerabilities we mentioned previously is that theory, or substantive knowledge, often lacks the requisite detail to completely guide model specification. When these gaps in theory result in a mispecified model, many researchers employ data-driven approaches to respecify or modify the model (MacCallum, Roznowski, & Necowitz, 1992). Machine learning offers new ways of building SEMs by learning directly from data. If theory generation is the domain of purely exploratory analyses, while theory appraisal is the traditional domain of SEM, then machine learning incorporates exploratory techniques into traditional SEM analyses to facilitate the expansion and fine-tuning of theory (Brandmaier & Jacobucci, 2021). Huang, Chen, and Weng (2017, p. 330) called machine learning SEM a “semi-confirmatory approach” because it can “embrace both existing theories and ambiguous relations that await further exploration”. In this vein, machine learning brings a shift in focus from explanation to prediction (Shmueli, 2010; Yarkoni & Westfall, 2017). Indeed, social science has historically been geared towards understanding the underlying causal mechanisms that govern social phenomena or human behavior. On the other hand, the goal of a machine learning analysis is usually to use data to improve prediction rather than to discover the “true model” (Berk, 2008; Brandmaier

& Jacobucci, 2021). Some have argued that applying machine learning to SEM marks a synthesis of model explanation and prediction in computational social science (e.g., Brandmaier & Jacobucci, 2021; Shmueli, 2010).

Regularized SEM, SEM Trees, and Forests—Two recently developed SEM methods that incorporate machine learning tools are regularized SEM, and SEM trees or forests. Regularization involves modifying a cost function to include a penalty term which either reduces the size of parameters or shrinks them to zero depending on the penalty. This technique, also called shrinkage or penalization, can reduce model complexity and overfitting, and overcome underidentification by allowing for small biases in prediction to decrease variance or improve efficiency. Regularization has been used for variable selection and model estimation in regression (Tibshirani, 1996), for dimensionality reduction (Zou, 2006), and in concert with graphical modeling tools (Friedman et al., 2008; for a broad overview, see Zou, 2006). Recently, regularization methods have also been incorporated into the SEM framework (e.g., Huang, 2020a; Huang et al., 2017; Jacobucci et al., 2019; Jacobucci et al., 2016). The regularized SEM has widespread applications. For instance, researchers use SEM with penalized ML estimation to perform semi-confirmatory factor analysis (Huang et al., 2017), variable selection in multiple cause multiple indicator models (Jacobucci et al., 2016), exploratory mediation analysis (Serang, Jacobucci, Brimhall, & Grimm, 2017), and model selection for individual dynamic models using time series data (Ye, Gates, Henry, & Luo, 2021).

SEM trees, first proposed by Brandmaier, von Oertzen, McArdle, and Lindenberger (2013), is a non-parametric method that learns hierarchical structures of observed data in the search for heterogeneous sets of parameter estimates in SEM. The decision tree algorithm identifies important predictors to recursively partition the complete dataset into maximally distinct subsets and finds the best-fitting SEM model in each resulting subgroup (Brandmaiser et al., 2013; Brandmaier & Jacobucci, 2021). SEM forests (Brandmaiser et al., 2016) are ensembles of SEM trees, each constructed with a random sample selected from the original data. Nonparametric estimates from a forest are more robust than those from single trees because they aggregate the predictive information across trees (Brandmaiser et al., 2016). The methods were first available in OpenMx and R packages (e.g., semtree; Brandmaiser et al., 2013), and have just recently been implemented in M*plus* (i.e., M*plus* Trees; Serang et al., 2021). SEM trees and forests have been combined with latent growth curve models to discover heterogenous growth trajectories (Usami, Jacobucci, & Hayes, 2019), applied to assessment data to predict students' attitudes towards collaborations (Li, Zhang, Li, Huang, & Shao, 2021) and applied to neuroscience data to study age differentiation in brain structures and change in cognitive functions across the life span (de Mooij, Henson, Waldorp, & Kievit, 2018).

Computation graphs and Deep Learning SEM for Complex Data Structures—Researchers are becoming increasingly interested in exploring more complex data structures using SEM, e.g., high-dimensional data produced by brain imaging, epigenetic or genomic measurements, and experimental models combining disparate data sources. For instance, analysts use SEM to investigate multi-trait genetic architecture (e.g., Pegolo et al., 2020) or to discover relationships between genetic variation and imaging variables in imaging genetics studies (e.g., Huisman et al., 2018). The newly developed Genomic SEM (Grotzinger et al., 2019) provides insights into the multivariate genetic architecture of complex traits. In these applications, the estimation of SEMs becomes even more challenging as latent variable models become higher dimensional and more complex.

In these modern higher-dimensional data applications, researchers seek to find ways to reformulate SEM using alternative, more flexible optimization algorithms to obtain parameter estimates. For example, van Kesteren and Oberski (2021) considered SEM as a *computation graph*, which enables the specification of a large variety of objective functions for SEM in different contexts. When analyzed as a computation graph, SEM researchers can take advantage of optimization routines well-studied in deep learning to estimate traditional SEM models. The integration with computation graphs and deep learning makes SEM an increasingly powerful tool for more complex data types and structures, which opens the door to applications in novel fields.

Statistical challenges in Machine Learning and Deep Learning SEM—The use of these methods raises questions about reproducibility and out-of-sample generalizability of the results. The risk of capitalizing on chance that accompanies the use of traditional exploratory SEM methods for variable selection (e.g., Lagrange multiplier test or modification indices; MacCallum, Roznowski, & Necowitz, 1992) also applies to the use of machine learning and deep learning SEM methods. For example, Yuan and Liu (2021) found that neither regularization nor the Lagrange multiplier test can reliably locate the true nonzero parameters or find the true model, even though regularized SEM is more robust to nonnormality regardless of the sample size. Another downside of common regularization methods is the instability of results and high false positive rates. Li and Jacobucci (2021) applied ideas from bootstrap aggregating (Breiman,1996) to incorporate stability selection in regularized SEM. However, the computational cost of these bootstrap procedures can be daunting.

One of the leading challenges facing machine learning and deep learning SEM researchers is a lack of methods for *post-model selection statistical inference*, which is also one of the primary sources for the reproducibility crisis that remains to be solved in statistics (e.g., Breiman, 1992; Leeb & Pötscher, 2006). SEM researchers rely on statistical hypotheses and inferences established under the assumption that the model is known in advance; however, the inference of a machine learning SEM with data-driven modeling departs from this approach. Naïve standard errors are inaccurate because both the randomness introduced by the model selection process and the sample space restriction implied by the chosen model must be accounted for (Huang, 2020b). However, making valid post selection inferences is a notoriously difficult task (Leeb & Pötscher, 2006). This active area of research (e.g., Huang, 2020b) will be key to the future development of machine learning and deep learning in SEM.

Time Series Analysis

Early on, analysts considered the use of SEM software to analyze covariance matrices from repeated measures data (e.g., Jöreskog, 1971, 1979). Even prior to this, Wright (1925) applied path analysis to time series data with latent variables. But it was not until the mid-1980s and later that traditional time series models were unified within the SEM framework (Hershberger, Corneal, & Molenaar, 1994; Hershberger, Molenaar, & Corneal, 1996; Molenaar, 1985, 1999; Van Buuren, 1997). Hamaker, Dolan and Molenaar (2002) clarified the nature of SEM-based estimates of traditional time series model parameters (e.g., autoregressive moving average models). Since the publication of these seminal works a rich tradition of modeling time series with SEM has emerged, capitalizing on the ability to specify measurement models for time series measured with error (Molenaar, 1985), impose and assess measurement invariance across time (Adolf, Schuurman, Borkenau, Borsboom & Dolan, 2014), handle categorical endogenous variables (Zhang & Nesselroade, 2007), utilize measures of model fit to quantify model error (Trichtinger & Zhang, 2021), and establish the relationship between similar modeling frameworks (e.g., SEM and state space modeling; MacCallum & Ashby, 1986; Chow, Ho, Hamaker & Dolan, 2010).

These works have become increasingly relevant over the last two decades as the availability of intensively collected repeated measures data has dramatically increased. This is due in part to technological advances, the ubiquity of smartphones and wearable technologies, and increasing interest in characterizing processes of change in social science research. Often these data are collected according to methods amenable to frequent measurements, such as daily diary methods, experience sampling methods, and ecological momentary assessment. The data characteristics associated with multivariate time series collected on multiple individuals present unique modeling challenges. To accommodate these new data types, as well as the increasing interest in dynamic modeling, several promising developments in the SEM and related frameworks have emerged. In this section we will briefly discuss three of these innovations and how they are being used to reshape the analysis of intensive longitudinal data.

Continuous-Time Modeling—Typically, approaches to time series analysis are classified according to whether time is treated as discrete or continuous. The former classification is discrete time, and the latter, continuous time. Continuous time models bring a few potential benefits to the study of change, including the ability to characterize lagged effects beyond the time intervals of data collection (Voelkle et al., 2012). This makes it easier to generalize results and relate findings from studies with different time intervals. Although discrete time approaches are more common in the social, health and behavioral sciences, continuous time approaches have become more popular as (a) more user-friendly software capable of fitting continuous time models becomes available, (b) interest in understanding behavior in the moment, often accomplished by alerting subjects and requesting a response at random intervals resulting in unequal time intervals, grows, and (c) the conceptual differences between the two approaches become better understood (Chow et al., In Press).

Analysts can estimate continuous time SEMs for linear time-dependent systems using a variety of different software platforms. In fact, researchers can use any SEM package capable of imposing constraints on the model implied moments and parameter matrices to fit basic (N=1) continuous time models. Recently, more fully featured software options have emerged for modeling dynamic systems with unequally spaced observations. These include Continuous Time Structural Equation Modeling (CTSEM; Driver, Oud & Voelkle, 2017) and dynr (Ou, Hunter & Chow, 2019), both freely available R packages with extensive documentation.

Dynamic Structural Equation Modeling—Another promising direction, particularly for synthesizing multiple subject time series data, is that of Dynamic Structural Equation Modeling (DSEM; Asparouhov et al., 2018). The DSEM modeling framework incorporates three distinct modeling traditions for extending many traditional time series models in interesting new directions. These traditions include (1) *time-series analysis*, a large number of repeated measurements are obtained for a single individual with the purpose of modeling lead-lag relations, (2) *multilevel modeling*, for analyzing multiple individuals simultaneously with the goal of characterizing and explaining quantitative individual differences in the parameters governing a dynamic process, and (3) *structural equation modeling*, allowing for the inclusion of latent variables, multiple-group analysis, testing measurement invariance, handling of missing data, categorical endogenous variables, model fit and many other features inherent to the SEM framework. The breadth of modeling tools available in the DSEM approach to handling time-ordered data represents a major contribution to the study of within-person change.

DSEM has been available since Mplus Version 8 (Muthen & Muthen, 2017). Importantly, DSEM models in Mplus are estimated in a Bayesian modeling framework using Markov Chain Monte Carlo (MCMC). Readers should note that although DSEM incorporates many features common to SEM analysis, the Bayesian statistical framework brings with it important differences in terms of fitting and interpreting SEM analyses. MCMC yields a distribution of possible values for unknown model parameters and common model fit indices (RMSEA, CFI, TLI) are not available. In addition, DSEM uses random effects to account for heterogeneity both between and within individuals. This brings implicit assumptions about how individuals differ in terms of change processes, and how those change processes evolve within an individual. Specifically, DSEM allows for time series model parameters to differ quantitatively (e.g., according to a multivariate normal distribution), limiting the amount of heterogeneity one is capable of accounting for in each analysis.

Group Iterative Multiple Model Estimation (GIMME)—A flexible approach within the SEM framework for analyzing multiple subject time series that overcomes some of the limitations of the random effects approach is Group Iterative Multiple Model Estimation (GIMME; Gates & Molenaar, 2012). In this approach, individual-level models have dynamics that are common to the group (and subgroup) and dynamics that are specific to each individual in the sample. Importantly, the GIMME algorithm uses modification indices (Sorbom, 1989) within a forward-selection specification search. The search begins with an empty model (or one with autoregressive parameters freely estimated) and terminates when at least two of four model fit indices indicate a close fit. The four fit measures in GIMME are the root mean square error of approximation (RMSEA; Steiger, 1990), standardized root mean square residual (SRMR; Bentler, 1995), comparative fit index (CFI; Bentler, 1990), and non-normed fit index (NNFI; Bentler and Bonett, 1980). GIMME uses RMSEA/SRMR values less than 0.05, and CFI/NNFI values greater than 0.95, to indicate a well-fitting model.

Since its inception, some interesting extensions have been developed to broaden the utility of the GIMME approach. These developments include the addition of subgroup analysis (Gates, Lane, et al., 2017), metrics to assess the robustness of subgroup solutions (Gates, Fisher, Arizmendi, et al.,2019), fMRI-specific extensions (Duffy et al., 2021), the inclusion of latent variables (Gates, Fisher & Bollen, 2021), and allowing contemporaneous relations to be more flexibly estimated (Luo et al., 2022). All these developments are also available in the freely distributed GIMME R package (Lane et al., 2021).

Summary of Future Developments in Time Series—In closing, the continuous time, DSEM, and GIMME frameworks are three recent examples of the continuing ebb and flow of diffusion and generalization associated with methodological development in SEM. As methodologists turn their attention towards time-dependent data with the aim of unifying and generalizing a large class of multiple-subject time series models, a new generation of researchers are exposed to SEM as a means to study and measure change. As technological developments make it easier to collect rich multivariate time series data, these trends towards the incorporation of new data types are only likely to accelerate.

CONCLUSION

In this article, we gave a brief history and overview of the evolution of SEM during the 50 years of *SSR*'s publication. Though the roots of SEM go back at least a century to the pioneering work of Wright (1918; 1921) and Spearman (1904), the pace of development was far more rapid in the last 50 years than before. The subtitle of our article summarizes the characteristic trends of this period as generalization, unification, and diffusion. The generalization aspect describes the sweeping nature of latent variable SEM that includes many of the most common statistical models applied in the social and behavioral sciences. From the simplicity of ANOVA to the complexity of multilevel or time series SEM with latent variables, the general model encompasses them. The unification aspect is related in that SEM provides a unifying framework where these diverse models are special cases of a more general model. Furthermore, the general model gives the power to create new hybrid models that combine features of seemingly distinct structures. SEM also unifies in that there are estimation procedures and fit statistics that accompany SEM. When we bring more specialized models into this framework, these estimators and fit statistics can provide novel methods of developing estimates or diagnosing model fit that might not be known in the context of the specialized models.

Diffusion is the third trend of SEM in the last 50 years that we highlighted. During this period, SEM spread through the social and behavioral sciences as well as into information science, computer science, ecology, and medicine, and even returned to its origins in genetics and biology. As it diffused, it acquired new characteristics. One of the most notable is the development of DAGs or NPSEM. Indeed, the integration of classic latent variable

SEM with NPSEM is ongoing. The diffusion of SEM through the disciplines we named is far from complete. Even within the same discipline, fields differ in their familiarity and use of SEM. Some fields are at the forefront while others have outdated or incorrect knowledge of SEM (Bollen & Pearl, 2013).

Our article documented the continuing role of SEM in *SSR* papers. There were too many such articles to describe each, but we illustrated diverse *SSR* applications. We also reflected on the strengths and vulnerabilities of SEM and active areas of development.

In brief, SEM is a widely used statistical modeling tool across many fields. It continues to evolve and to become more general. *SSR* has been home to many SEM applications over its 50 years. We anticipate more to come.

Acknowledgments

We gratefully acknowledge partial support for this work from NIH (1R21MH119572-01, P2C HD050924, and T32 HD091058).

REFERENCES

Adolf J, Schuurman NK, Borkenau P, Borsboom D, & Dolan CV (2014). Measurement invariance within and between individuals: A distinct problem in testing the equivalence of intra-and inter-individual model structures. Frontiers in Psychology, 5.

Allison PD, & Bollen KA (1997, August 9–13). Change score, fixed effects and random component models: A structural equation approach [Paper presentation]. ASA 92nd Annual Meeting, Toronto, Canada.

Alwin DF, & Hauser RM (1975). The decomposition of effects in path analysis. American Sociological Review, 40(1), 37–47.

Angist JD, & Pischke J-S (2015). Mastering ‘metrics: The path from cause to effect. Princeton University Press.

Asparouhov T, Hamaker EL, & Muthén B. (2018). Dynamic Structural Equation Models, Structural Equation Modeling, 25:3, 359–388, DOI: 10.1080/10705511.2017.1406803

Asparouhov T, & Muthén B. (2009). Exploratory structural equation modeling. Structural, Equation Modeling, 16, 397–438.

Athey S, Imbens GW, & Wager S. (2018). Approximate residual balancing: Debiased, inference of average treatment effects in high dimensions. Journal of the Royal Statistical, Society. Series B, Statistical Methodology, 80, 597–623 10.1111/rssb.12268

Baron RM, & Kenny DA (1986). The moderator–mediator variable distinction in social, psychological research: Conceptual, strategic, and statistical considerations. Journal of, Personality and Social Psychology, 51(6), 1173. [PubMed: 3806354]

Barrett P. (2007). Structural equation modeling: Adjudging model fit. Personality and, Individual Differences, 42(5), 815–824.

Bauer DJ, & Curran PJ (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. Psychological Methods, 9(1), 3. [PubMed: 15053717]

Beierlein C, Kuntz A, & Davidov E. (2016). Universalism, conservation, and attitudes toward, minority groups. Social Science Research, 58, 68–79. [PubMed: 27194652]

Bentler PM (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107(2), 238–246. [PubMed: 2320703]

Bentler PM (1995). EQS structural equations program manual (Ver. 6). Multivariate, Software.

Bentler PM, & Bonett DG (1980). Significance tests and goodness of fit in the analysis of, covariance structures. Psychological Bulletin, 88(3), 588.

Berk RA (2008). Statistical learning from a regression perspective (Springer Texts in, Statistics, Vol. 14). Springer.

Bianconcini S, & Bollen KA (2018). The latent variable-autoregressive latent trajectory, model: A general framework for longitudinal data analysis. Structural Equation, Modeling, 25(5), 791–808. [PubMed: 31293345]

Blalock HM (1961). Correlation and causality: The multivariate case. Social Forces, 39(3), 246–251. 10.2307/2573216

Blalock HM (1963). Making causal inferences for unmeasured variables from correlations, among indicators. American Journal of Sociology, 69(1), 53–62.

Blalock HM (1964) [1961]. Causal inferences in nonexperimental research. The University of, North Carolina Press.

Blalock HM (Ed.). (1971). Causal models in the social sciences. Aldine Publishing Company.

Boggess LN (2017). Disentangling the reciprocal relationship between change in crime and, racial/ethnic change. Social Science Research, 68, 30–42. [PubMed: 29108598]

Bollen KA (1987). Total, direct, and indirect effects in structural equation models. Sociological Methodology, 17, 37–69.

Bollen KA (1989a). Structural equations with latent variables. John Wiley & Sons.

Bollen KA (1989b). A new incremental fit index for general structural equation models. Sociological Methods & Research, 17(3), 303–316. 10.1177/2F0049124189017003004

Bollen KA (1995). Structural equation models that are nonlinear in latent variables: A least-squares estimator. Sociological Methodology, 25, 223–251.

Bollen KA (1996). An alternative two stage least squares (2SLS) estimator for latent variable, equations. Psychometrika, 61(1), 109–121. 10.1007/BF02296961

Bollen KA (2002). Latent variables in psychology and the social sciences. Annual Review of, Psychology, 53(1), 605–634.

Bollen KA (2019). Model implied instrumental variables (MIIVs): An alternative orientation, to structural equation modeling. Multivariate Behavioral Research, 54(1), 31–46. 10.1080/00273171.2018.1483224 [PubMed: 30222004]

Bollen KA, & Brand JE (2010). A general panel model with random and fixed effects: A, structural equations approach. Social Forces, 89(1), 1–34.

Bollen KA, & Curran PJ (2006). Latent curve models: A structural equation perspective. John Wiley & Sons.

Bollen KA, & Diamantopoulos A. (2017). In defense of causal-formative indicators: A minority report. Psychological Methods, 22(3), 581. [PubMed: 26390170]

Bollen KA, Fisher ZF, Giordano ML, Lilly AG, Luo L, & Ye A. (2021). An introduction to model implied instrumental variables using two stage least squares (MIIV-2SLS) in structural equation models (SEMs). Psychological Methods.

Bollen KA, Gutin I, Halpern CT, & Harris KM (2021). Subjective health in adolescence: Comparing the reliability of contemporaneous, retrospective, and proxy reports of overall health. Social Science Research, 96, 102538.

Bollen K, & Lennox R. (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110(2), 305.

Bollen KA, & Pearl J. (2013). Eight myths about causality and structural equation models. In Handbook of causal analysis for social research (pp. 301–328). Springer, Dordrecht.

Bollen KA, & Ting K. (1993). Confirmatory tetrad analysis. Sociological Methodology, 23, 147–175.

Borsboom D, & Cramer AO (2013). Network analysis: an integrative approach to the structure of psychopathology. Annual Review of Clinical Psychology, 9, 91–121.

Borsboom D, Mellenbergh GJ, & van Heerden J. (2003). The theoretical status of latent variables. Psychological Review, 110(2), 203. [PubMed: 12747522]

Brandmaier A. & Jacobucci R. (2021). Machine-Learning Approaches to Structural Equation Modeling. Handbook of SEM (2nd Edition). Guilford.

Brandmaier AM, von Oertzen T, McArdle JJ, & Lindenberger U. (2013). Structural equation model trees. Psychological Methods, 18(1), 71–86. 10.1037/a0030001 [PubMed: 22984789]

Brandmaier AM, Prindle JJ, McArdle JJ, & Lindenberger U. (2016). Theory-guided exploration with structural equation model forests. Psychological Methods 21, 566–582. doi: 10.1037/met0000090 [PubMed: 27918182]

Branic N, & Hipp JR (2018). Growing pains or appreciable gains? Latent classes of neighborhood change, and consequences for crime in Southern California neighborhoods. Social Science Research, 76, 77–91. [PubMed: 30268285]

Breiman L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. Journal of the American Statistical Association, 87(419), 738–754. doi: 10.1080/01621459.1992.10475276

Breiman L. (1996). Bagging predictors. Machine Learning 24, 123–140 10.1007/BF00058655

Breiman L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Browne MW, MacCallum RC, Kim CT, Andersen BL, & Glaser R. (2002). When fit indices and residuals are incompatible. Psychological Methods, 7(4), 403. [PubMed: 12530701]

Cartwright N. (1999). The dappled world: A study of the boundaries of science. Cambridge University Press.

Chen F, Curran PJ, Bollen KA, Kirby J, & Paxton P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. Sociological Methods & Research, 36(4), 462–494. [PubMed: 19756246]

Chow S-M, Losardo D, Park J, & Molenaar PCM (In Press). Continuous-Time Dynamic Models: Connections to Structural Equation Models and Other Discrete-Time Models. In Hoyle RH, Structural Equation Modeling: Concepts, Issues, and Applications (2nd Ed.). New York: Guilford.

Chow S-M, Ho MR, Hamaker EL, & Dolan CV (2010). Equivalence and differences between structural equation modeling and state-space modeling techniques. Structural Equation Modeling, 17(2), 303–332.

Cliff N. (1983). Some cautions concerning the application of causal modeling methods. Multivariate Behavioral Research, 18, 115–126. 10.1207/s15327906mbr1801_7 [PubMed: 26764558]

Cook TD, Campbell DT, & Day A. (1979). Quasi-experimentation: Design & analysis issues for field settings (Vol. 351). Boston: Houghton Mifflin.

de Mooij S, Henson R, Waldorp LJ, & Kievit RA (2018). Age Differentiation within Gray Matter, White Matter, and between Memory and White Matter in an Adult Life Span Cohort. The Journal of Neuroscience, 38(25), 5826–5836. 10.1523/JNEUROSCI.1627-17.2018 [PubMed: 29848485]

DeYoung CG, & Krueger RF (2020). To wish impossible things: On the ontological status of latent variables and the prospects for theory in psychology. Psychological Inquiry, 31(4), 289–296.

Driver CC, Oud JHL, Voelkle MC (2017). “Continuous Time Structural Equation Modeling with R Package ctsem.” Journal of Statistical Software, 77(5), 1–35.

Duncan OD (1966). Path analysis: Sociological examples. American Journal of Sociology, 72(1), 1–16.

Duncan OD (1975). Introduction to structural equation models. Academic Press.

Duffy KA, Fisher ZF, Arizmendi CA, Molenaar PCM, Hopfinger J, Cohen JR, Beltz AM, Lindquist MA, Hallquist MN, & Gates KM (2021). Detecting Task-Dependent Functional Connectivity in Group Iterative Multiple Model Estimation with Person-Specific Hemodynamic Response Functions. Brain Connectivity, 11(6), 418–429. [PubMed: 33478367]

Elwert F, & Winship C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. Annual review of sociology, 40, 31.

Eronen MI, & Bringmann LF (2021). The theory crisis in psychology: How to move forward. Perspectives on Psychological Science, 1745691620970586.

Fornell C, & Larcker DF (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50.

Fried EI (2020). Lack of theory building and testing impedes progress in the factor and network literature. Psychological Inquiry, 31(4), 271–288. 10.1080/1047840X.2020.1853461

Friedman J, Hastie T, & Tibshirani R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3), 432–441. [PubMed: 18079126]

Gates KM & Molenaar PCM (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. NeuroImage, 63, 310–319. [PubMed: 22732562]

Gates KM, Fisher ZF, Arizmendi C, Henry TR, Duffy KA, & Mucha PJ (2019). Assessing the robustness of cluster solutions obtained from sparse count matrices. Psychological Methods, 24(6), 675. [PubMed: 30742473]

Gates KM, Lane ST, Varangis E, Giovanello K, & Guskiewicz K. (2017). Unsupervised classification during time-series model building. Multivariate Behavioral Research, 52(2), 129–148. [PubMed: 27925768]

Gates KM, Fisher ZF, & Bollen KA (2020). Latent variable GIMME using model implied instrumental variables (MIIVs). Psychological methods, 25(2), 227–242. 10.1037/met0000229 [PubMed: 31246041]

Goldberger AS & Duncan OD (Eds.). (1973). Structural equation models in the social sciences. Academic Press.

Grotzinger AD, Rhemtulla M, de Vlaming R. (2019). Genomic structural equation modeling provides insights into the multivariate genetic architecture of complex traits. Nat Hum Behav 3, 513–525. 10.1038/s41562-019-0566-x [PubMed: 30962613]

Hamaker EL, Dolan CV, & Molenaar PCM (2002). On the Nature of SEM Estimates of ARMA Parameters. Structural Equation Modeling, 9(3), 347–368.

Hayduk LA (2014). Shame for disrespecting evidence: the personal consequences of insufficient respect for structural equation model testing. BMC Medical Research Methodology, 14(1), 1–10. [PubMed: 24383436]

Hayduk L, Cummings G, Boadu K, Pazderka-Robinson H, & Boulianne S. (2007). Testing! testing! one, two, three–Testing the theory in structural equation models!. Personality and Individual Differences, 42(5), 841–850.

Heckman JJ (2005). The scientific model of causality. Sociological methodology, 35(1), 1–97.

Hershberger SL (1994). The specification of equivalent models before the collection of data. In von Eye A. and Clogg CC (Eds.), Latent variables analysis (pp. 68–108). Thousand Oaks, CA: Sage.

Hershberger SL, Corneal SE, & Molenaar PCM (1994). Dynamic factor analysis: An application to emotional response patterns underlying daughter/father and stepdaughter/stepfather relationships. Structural Equation Modeling, 2, 31–52.

Hershberger SL, Molenaar PCM, & Corneal SE (1996). A hierarchy of univariate and multivariate structural time series models. In Marcoulides GA & Schumacker RE (Eds.), Advanced structural equation modeling: Issues and techniques (pp. 159–194). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Holland PW (1986). Statistics and causal inference. Journal of the American Statistical Association, 81(396), 945–960.

Hoyle RH, & Panter AT (1995). Writing about structural equation models. In Hoyle RH(Ed.), Structural equation modeling: Concepts, issues, and applications (pp. 158–176). Sage Publications.

Hu LT, & Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55.

Huang PH (2020a). lslx: Semi-confirmatory structural equation modeling via penalized likelihood. Journal of Statistical Software, 93, 1–37.

Huang PH (2020b) Postselection Inference in Structural Equation Modeling, Multivariate Behavioral Research,55(3), 344–360, DOI: 10.1080/00273171.2019.1634996 [PubMed: 31304796]

Huang PH, Chen H, & Weng L-J (2017). A penalized likelihood method for structural equation modeling. Psychometrika, 82(2), 329–354. [PubMed: 28417228]

Huisman SMH, Mahfouz A, Batmanghelich NK (2018). A structural equation model for imaging genetics using spatial transcriptomics. Brain Inf. 5, 13. 10.1186/s40708-018-0091-0

Jacobucci R, Brandmaier AM, & Kievit RA (2019). A Practical Guide to Variable Selection in Structural Equation Modeling by Using Regularized Multiple-Indicators, Multiple-Causes Models. Advances in Methods and Practices in Psychological Science, 55–76. 10.1177/2515245919826527 [PubMed: 31463424]

Jacobucci R, Grimm KJ, & McArdle JJ (2016). Regularized structural equation modeling. Structural Equation Modeling, 23(4), 555–566. [PubMed: 27398019]

Jöreskog KG (1971). Estimation and testing of simplex models. The British Journal of Mathematical and Statistical Psychology, 23(2), 121–145.

Jöreskog KG (1973). A general method for estimating a linear structural equation system. In Goldberger S. & Duncan OD (Eds.), Structural equation models in the social sciences (pp. 85–112). Academic Press.

Jöreskog KG (1979). Statistical estimation of structural models in longitudinal-developmental investigations. In Nesselroade JR & Baltus PB (Eds.), Longitudinal research in the study of behavior and development (pp. 303–351). New York: Academic.

Jöreskog KG & Sörbom D. (1974). LISREL III [Computer software].

Jöreskog KG & Sörbom D. (1981). LISREL. Analysis of Linear Structural Relationships by the Method of Maximum Likelihood. Version V User's Guide. National Educational Resources.

Jöreskog KG & Sörbom D. (1985). PRELIS: A Program for Multivariate Data Screening and Data Summarization: A Preprocessor for LISREL. University of Uppsala, Sweden.

Kaplan D, & Depaoli S. (2012). Bayesian structural equation modeling. In Hoyle RH(Ed.), Handbook of structural equation modeling (pp. 650–673). The Guilford Press.

Keesling JW (1973). Maximum likelihood approaches to causal flow analysis (Publication No. 302695981) [Doctoral dissertation, The University of Chicago]. ProQuest Dissertations & Theses Global.

Kenny DA (1979). Correlation and causality. Wiley.

Kenny DA, & McCoach DB (2003). Effect of the number of variables on measures of fit in structural equation modeling. Structural equation modeling, 10(3), 333–351.

Kim ES, Cao C, Wang Y, & Nguyen DT (2017). Measurement invariance testing with many groups: A comparison of five approaches. Structural Equation Modeling, 24(4), 524–544. 10.1080/10705511.2017.1304822

Kirby JB, & Bollen KA (2009). Using instrumental variable tests to evaluate model specification in latent variable structural equation models. Sociological Methodology, 39(1), 327–355. 10.1111/j.1467-9531.2009.01217.x [PubMed: 20419054]

Klein SB (2014). What can recent replication failures tell us about the theoretical commitments of psychology?. Theory & Psychology, 24(3), 326–338.

Land KC (1973). Identification, Parameter Estimation, and Hypothesis Testing in Recursive Sociological Models. In Goldberger AS & Duncan OD (Eds.), Structural equation models in the social sciences (pp. 19–49). Academic Press.

Lane S, Gates K, Fisher Z, Arizmendi C, Molenaar P, Hallquist M, Pike H, Henry T, Duffy K, Luo L, Beltz A, Wright A, Park J, & Alvarez SC (2021). gimme: Group iterative multiple model estimation (R package version 0.7–7) [Computer software]. The Comprehensive R Archive Network. Available from https://github.com/GatesLab/gimme/

Lee S, & Hershberger S. (1990). A simple rule for generating equivalent models in covariance structure modeling. Multivariate Behavioral Research, 25, 313–334. [PubMed: 26761406]

Leeb H, & Pötscher BM (2006). Can one estimate the conditional distribution of post-model-selection estimators? The Annals of Statistics, 34(5), 2554–2591. doi: 10.1214/009053606000000821

Li J, Zhang M, Li Y, Huang F, & Shao W. (2021). Predicting Students' Attitudes toward collaboration: evidence from structural equation model trees and forests. Front. Psychol 12:604291. doi: 10.3389/fpsyg.2021.604291 [PubMed: 33841240]

Li X, & Jacobucci R. (2021). Regularized structural equation modeling with stability selection. Psychological methods. Advance online publication. 10.1037/met0000389

Lord FM (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13(4), 517–549.

Lord FM, & Novick MR (2008). Statistical theories of mental test scores. IAP.

Luo L, Fisher ZF, Arizmendi C, Molenaar P, Beltz A, & Gates KM (2022). Estimating both directed and undirected contemporaneous relations in time series data using hybrid-group iterative multiple model estimation. Psychological Methods.

MacCallum RC (1990). The need for alternative measures of fit in covariance structure modeling. Multivariate Behavioral Research, 25(2), 157–162. [PubMed: 26794477]

MacCallum RC, & Ashby GF (1986). Relationships between linear systems theory and covariance structure modeling. Journal of Mathematical Psychology, 30(1), 1–27.

MacCallum RC, Roznowski M, & Necowitz LB (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. Psychological Bulletin, 111(3), 490–504. [PubMed: 16250105]

MacCallum RC, Wegener DT, Uchino BN, & Fabrigar LR (1993). The problem of equivalent models in applications of covariance structure analysis. Psychological Bulletin, 114(1), 185. [PubMed: 8346326]

Marsh HW, Hau KT, & Wen Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural Equation Modeling, 11(3), 320–341.

McCutcheon AC (1987). Latent class analysis. Beverly Hills, CA: Sage.

McNeish D, An J, & Hancock GR (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. Journal of Personality Assessment, 100(1), 43–52. [PubMed: 28631976]

McNeish D, & Wolf MG (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. Psychological Methods. Advance online publication. https://psycnet.apa.org/doi/10.1037/met0000425

Meredith W, & Tisak J. (1990). Latent curve analysis. Psychometrika, 55(1), 107–122.

Meuleman B, Roosma F, & Abts K. (2020). Welfare deservingness opinions from heuristic to measurable concept: The CARIN deservingness principles scale. Social Science Research, 85, 102352.

Molenaar PCM (1985). A dynamic factor model for the analysis of multivariate time series. Psychometrika, 50, 181–202.

Molenaar PCM (1999). Comment on fitting MA time series by structural equation models. Psychometrika, 64, 91–94.

Morgan SL, & Winship C. (2015). Counterfactuals and causal inference. Cambridge University Press.

Muthén B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49(1), 115–132.

Muthén BO (2011). Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. Retrieved from http://www.statmodel2.com/download/causalmediation.pdf

Muthén B. & Asparouhov T. (2013). BSEM measurement invariance analysis. Mplus Web Notes: No. 17. https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén LK and Muthén BO (2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Muthukrishna M, & Henrich J. (2019). A problem in theory. Nature Human Behaviour, 3(3), 221–229.

Nunally JC, & Bernstein I. (1978). Psychometric Theory. New York McGraw.

Oberauer K, & Lewandowsky S. (2019). Addressing the theory crisis in psychology. Psychonomic Bulletin & Review, 26(5), 1596–1618. [PubMed: 31515732]

Ou L, Hunter MD, Chow S-M (2019). "What's for dynr: A package for linear and nonlinear dynamic modeling in R." The R Journal, 11, 1–20.

Pearl J. (2000). Causality: Models, reasoning, and inference. Cambridge University Press.

Pearl J. (2012). The causal mediation formula–A guide to the assessment of pathways and mechanisms. Prevention Science, 13, 426–436. 10.1007/s11121-011-02701 [PubMed: 22419385]

Pearl J, & Mackenzie D. (2018). The book of why: the new science of cause and effect. Basic books.

Pegolo S, Momen M, & Morota G. (2020). Structural equation modeling for investigating multi-trait genetic architecture of udder health in dairy cattle. Scientific Reports, 10, 7751. 10.1038/s41598-020-64575-3 [PubMed: 32385377]

Perreira KM, Deeb-Sossa N, Harris KM, & Bollen KB (2005). What are we measuring? An evaluation of the CES-D across race/ethnicity and immigrant generation. Social Forces, 83(4), 1567–1601. 10.1353/sof.2005.0077

Pudrovska T, Logan ES, & Richman A. (2014). Early-life social origins of later-life body weight: the role of socioeconomic status and health behaviors over the life course. Social Science Research, 46, 59–71. [PubMed: 24767590]

Radloff LS (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 1(3), 385–401. 10.1177/014662167700100306

Raftery AE (1995). Bayesian model selection in social research. Sociological Methodology, 25, 111–163.

Raudenská P. (2020). The cross-country and cross-time measurement invariance of positive and negative affect scales: Evidence from European Social Survey. Social Science Research, 86, Article 102369. 10.1016/j.ssresearch.2019.102369

Raykov T, & Marcoulides GA (2001). Can there be infinitely many models equivalent to a given covariance structure model?. Structural Equation Modeling, 8(1), 142–149.

Raykov T, & Penev S. (1999). On structural equation model equivalence. Multivariate Behavioral Research, 34, 199–244. [PubMed: 26753936]

Rigdon EE (1994). Demonstrating the effects of unmodeled random measurement error. Structural Equation Modeling, 1(4), 375–380.

Rogosa D. (1987). Casual models do not support scientific conclusions: A comment in support of Freedman. Journal of Educational Statistics, 12(2), 185–195.

Roos JM, & Bauldry S. (2022). Confirmatory Factor Analysis. Sage Publishing.

Rosenbaum PR, & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.

Serang S, Jacobucci R, Brimhall KC, & Grimm KJ (2017) Exploratory Mediation Analysis via Regularization, Structural Equation Modeling, 24:5, 733–744, DOI: 10.1080/10705511.2017.1311775 [PubMed: 29225454]

Serang S, Jacobucci R, Stegmann G, Brandmaier AM, Culianos D, & Grimm KJ (2021). Mplus Trees: Structural Equation Model Trees Using Mplus, Structural Equation Modeling, 28:1, 127–137, DOI: 10.1080/10705511.2020.1726179

Seddig D, & Lomazzi V. (2019). Using cultural and structural indicators to explain measurement noninvariance in gender role attitudes with multilevel structural equation modeling. Social Science Research, 84, 102328. [PubMed: 31674334]

Shen W, Hu LC, & Hannum E. (2021). Effect pathways of informal family separation on children’s outcomes: Paternal labor migration and long-term educational attainment of left-behind children in rural China. Social Science Research, 97, 102576.

Shmueli G. (2010). To explain or to predict? Statistical Science, 25(3), 289–310.

Silvia ESM, & MacCallum RC (1988). Some factors affecting the success of specification searches in covariance structure modeling. Multivariate Behavioral Research, 23, 297–326. 10.1207/s15327906mbr2303_2 [PubMed: 26776526]

Sobel ME (1982). Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology, 13, 290–312.

Sorbom D. (1989). Model modification. Psychometrika, 54, 371–384.

Spearman C. (1904). “General intelligence,” objectively determined and measured. The American Journal of Psychology, 15(2), 201–292. https://www.jstor.org/stable/1412107

Steiger JH (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. Multivariate Behavioral Research, 25(2), 173–180. [PubMed: 26794479]

Steiger JH (2016). Notes on the Steiger-Lind (1980) handout. Structural Equation Modeling, 23(6), 777–781. 10.1080/10705511.2016.1217487

Steiger JH & Lind JC (1980, May). Statistically-based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society. Iowa City, IA.

Stelzl I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. Multivariate Behavioral Research, 21, 309–331. [PubMed: 26800964]

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58 (1), 267–288.

Trichtinger LA, & Zhang G. (2021). Quantifying Model Error in P-technique Factor Analysis. Multivariate Behavioral Research, 56(1), 41–56. [PubMed: 32000534]

Tucker LR, & Lewis C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, (1–10). 10.1007/BF02291170

Usami S, Jacobucci R. & Hayes T. (2019). The performance of latent growth curve model-based structural equation model trees to uncover population heterogeneity in growth trajectories. Computational Statistics, 34, 1–22. https://doi-org.libproxy.lib.unc.edu/10.1007/s00180-018-0815-x

van Bork R, Rhemtulla M, Waldorp LJ, Kruis J, Rezvanifar S, & Borsboom D. (2021). Latent variable models and networks: Statistical equivalence and testability. Multivariate Behavioral Research, 56(2), 175–198. [PubMed: 31617420]

Van Buuren S. (1997). Fitting ARMA time series by structural equation models. Psychometrika, 62, 215–236.

VanderWeele T. (2012). Invited commentary: Structural equation models and epidemiologic analysis, American Journal of Epidemiology, 176(7), 608–612, 10.1093/aje/kws213 [PubMed: 22956513]

van Kesteren E-J, & Oberski DL (2021). Flexible extensions to structural equation models using computation graphs. Structural Equation Modeling, 10.1080/10705511.2021.19715273

Vernon T, & Eysenck S. (Eds.). (2007). Special issue on structural equation modeling [Special issue]. Personality and Individual Differences, 42(5).

Voelkle M, Oud JHL, Davidov E, & Schmidt P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. Psychological Methods, 17, 176–192. [PubMed: 22486576]

Wallace D. (2012). Examining fear and stress as mediators between disorder perceptions and personal health, depression, and anxiety. Social Science Research, 41(6), 1515–1528. [PubMed: 23017971]

Wiley DE (1973). The identification problems for structural equation models with unmeasured variables. In Goldberger AS & Duncan OD (Eds.), Structural equation models in the social sciences (pp. 69–83). Academic Press.

Winship C, & Mare RD (1983). Structural equations and path analysis for discrete data. American Journal of Sociology, 89(1), 54–110.

Wright PG (1928). The tariff on animal and vegetable oils. The Macmillan Company.

Wright S. (1918). On the nature of size factors. Genetics, 3, 367–374. [PubMed: 17245910]

Wright S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. Proceedings of the National Academy of Sciences of the United States of America, 6(6), 320. [PubMed: 16576506]

Wright S. (1921). Correlation and causation. Journal of Agricultural Research, 20(7), 557–585.

Wright S. (1925). Corn and hog correlations. (Department Bulletin No. 1300). United States Department of Agriculture.

Wright S. (1934). The method of path coefficients. The Annals of Mathematical Statistics, 5(3), 161–215.

Yarkoni T, & Westfall J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100–1122. [PubMed: 28841086]

Ye A, Gates KM, Henry TR, & Luo L. (2021). Path and directionality discovery in individual dynamic models: A regularized unified structural equation modeling approach for hybrid vector autoregression. Psychometrika, 86, 404–441. [PubMed: 33840003]

Yuan K. & Liu F. (2021). Which Method is More Reliable in Performing Model Modification: Lasso Regularization or Lagrange Multiplier Test? Structural Equation Modeling, 28:1, 69–81, DOI: 10.1080/10705511.2020.1768858

Yucel D, & Borgmann LS (2022). Work–family conflict and depressive symptoms among dual-earner couples in Germany: A dyadic and longitudinal analysis. Social Science Research, 102684.

Zhang Z, & Nesselroade JR (2007). Bayesian Estimation of Categorical Dynamic Factor Models. Multivariate Behavioral Research, 42(4), 729–756.

Zou H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418–1429.