Overview

Structural equation modeling (SEM), also known as latent variable modeling, refers to a class of statistical techniques for investigating interrelationships among variables (Bollen, 1989). Variables can be continuous or categorical, and can be directly observable or not directly observable (i.e., latent). Factor analysis, path analysis, and the structural regression model are three basic types of SEM. Many traditional statistical procedures such as t-test, analysis of variance (ANOVA), and multiple regression analysis, can be conceptualized as special types of SEM. Historical foundations of SEM may date back to the late 1880s and early 1900s. In the past few decades, SEM has received much attention from both methodologists and practitioners and has been applied in many research fields including second language acquisition and individual differences (IDs). In this chapter, we briefly describe the overall procedure of conducting SEM. We then explain some technical features of typical SEMs, focusing on structural regression models and factor analysis, followed by a summary of SEM applications in ID research. Readers are assumed to be familiar with concepts and principles of regression analysis, and are referred to Kline (2016) and Kaplan (2009) for an introduction to SEM and Bollen (1989) for a more technical treatment of SEM.

An important distinction of SEM from classical statistical techniques that have been dominant in ID research (e.g., regression analysis) is that SEM allows for the inclusion of latent variables. Latent variables represent constructs or concepts under researchers’ investigation, for example, L2 learners’ motivation, self-regulation, personality traits, and aptitudes. Such constructs are not directly observable from individuals; instead, they are inferred from manifest variables (i.e., measurement indicators). Individuals’ responses to measurement indicators should be correlated with each other because they share a common cause: the latent variable. As such, this type of measurement indicator is called a reflective indicator, and latent variables (latent factors) are defined on the basis of correlations among measurement indicators. When using classical statistical procedures such as multiple regression analysis or ANOVA, a composite variable (e.g., by summing or averaging scores across measurement indicators) is often used to approximate the latent construct. A composite score contains two components: a true score and an error score. Therefore, results based on composite scores tend to be contaminated by measurement errors. SEM allows researchers to investigate the relationships among latent variables that are free from measurement errors. We illustrate the use of composite scores in SEM while controlling for measurement errors in a later section.
SEM and Factor Analysis

A typical SEM application involves five steps: 1) model specification, 2) model identification, 3) model estimation, 4) model evaluation (and model re-specification when needed) and model comparison, and 5) result reports and interpretations. These steps are well described in introductory SEM textbooks (e.g., Kline, 2016). Below, we briefly summarize the step procedure and discuss a few technical aspects of SEM that are likely to be encountered by ID researchers. To facilitate discussion without involving extensive equations and matrix operations, we use a hypothetical example in ID research.

Technical Features

In the running example, a researcher is interested in testing the effects of shame and guilt on L2 learners' motivation and language achievement. The researcher hypothesizes that shame and guilt have a differential direct effect on L2 learners' motivation, which in turn impacts L2 learners' language achievement. Shame, guilt, motivation, and achievement are latent constructs. The researcher adopts existing instruments with sound psychometric properties to measure these four constructs. The hypothesized relationships among the four latent constructs are depicted in the path diagram in Figure 32.1, in which shame ($F_1$), guilt ($F_2$), and motivation ($F_3$) are each measured by ten measurement indicators ($V_{11}-V_{20}$, $V_{11}-V_{20}$, and $V_{21}-V_{30}$, respectively) and achievement is measured by three indicators ($V_{31}-V_{33}$). In the path diagram, the four latent variables are represented by circles, and measurement indicators are represented by rectangles.

Overall Procedure for Conducting SEM

A structural equation model is generally theory driven. The hypothesized relationships specified in the running example should be derived from theories or previous studies, which constitute the first step of SEM: model specification. In this step, it is important to distinguish between two types of variables: exogenous versus endogenous. Exogenous variables are those with their causes being unknown or not explicitly specified in the model, while endogenous variables are those impacted by other variables in the model. Exogenous and endogenous variables are similar to predictors and outcome variables in multiple regression, respectively. However, an added feature of SEM over
conventional regression models is that an endogenous variable may serve as both a predictor and
an outcome (called intervening variable and sometimes mediator). In the example, shame and guilt
are exogenous because they both predict motivation. Motivation, achievement, and all 33 measure-
ment indicators are endogenous because they are predicted by other variables. Motivation is an
intervening variable because it is predicted by shame and guilt and also predicts achievement. The
distinction between exogenous and endogenous variables is closely related to the other crucial
task in the model specification step, that is, determination of directionality of relationships among
variables. In the example, shame and guilt are hypothesized to impact motivation and achievement,
but not the other way, based on researchers’ understanding of theories. Each directional relation-
ship is indicated by a single-headed arrow pointing from a variable to an endogenous variable in
a path diagram. A double-headed curve represents the variance of an exogenous variable (e.g.,
the variance of shame), the variance of a residual (e.g., the variance of measurement error for V_i),
the covariance between two exogenous variables (e.g., the covariance between shame and guilt),
or the covariance between two residuals (the hypothesized model does not contain any residual
covariance). One purpose of SEM analysis is to test whether such a specification is supported by
the data at hand.

This model is a structural regression model containing both a measurement part and a structural
part. The former captures directional relationships among the four latent factors and 33 measure-
ment indicators, and the latter captures relationships among the factors. These directional relation-
ships can be expressed as a set of regression equations much like those in multiple regression. The
structural part includes two equations:

\[
F_3 = \beta_{01} + \beta_1 F_1 + \beta_2 F_2 + D_1
\]

\[
F_4 = \beta_{02} + \beta_3 F_3 + D_2,
\]

(32.1)

where \(\beta_{01}\) and \(\beta_{02}\) denote the intercepts for \(F_3\) and \(F_4\), respectively. They are the expected values of \(F_3\) and \(F_4\) when the predicting factors have scores of zero. \(D_1\) and \(D_2\) are the residuals for \(F_3\) and \(F_4\) (often called disturbance), respectively. Residuals represent the scores on \(F_3\) and \(F_4\) that are not explained by the predicting factors. \(\beta_1\) to \(\beta_3\) are path coefficients quantifying directional relationships among latent factors. Path coefficients are interpreted in the same way as slope coefficients in multiple regression, representing the unique contribution of a predicting variable to an endog-
eneous variable. The measurement part includes 33 equations, one for each indicator. An efficient
way is to use vectors to express 33 equations in one equation. Here we present the equation only
for the first indicator \((V_1)\) measuring shame \((F_1)\).

\[
V_1 = \tau_1 + \lambda_1 F_1 + \epsilon_1,
\]

(32.2)

where \(\tau_1\) denotes the intercept for the first indicator, \(\lambda_1\) is the factor loading quantifying the lin-
ear relationship between \(V_1\) and \(F_1\), and \(\epsilon_1\) denotes the residual for this indicator which is often
conceptualized as measurement error. Residuals are assumed to be independently and normally
distributed with a mean of zero, and are uncorrelated with factors. All equations are analyzed
simultaneously, and thus SEM is sometimes called simultaneous equation modeling. Some types of
SEM can be conducted based on only variances and covariances among observed variables (e.g.,
path analysis); thus, such models are also called covariance structure models. Variances and covari-
ances among the observed variables form the observed covariance matrix. Other types require
also means of variables (e.g., longitudinal growth curve model) and are therefore also referred
to as mean and covariance structure models. In most SEM software such as AMOS (Arbuckle,
2019), EQS (Bentler & Wu, 2006), the lavaan package in R (Rosseel, 2012), and Mplus (Muthén
& Muthén, 1998–2019), the default setting is to include both covariance and mean structures
when raw data are provided.
The second step of SEM is to determine whether the hypothesized model is identified. Model identification should be determined prior to data collection. For a covariance structure model, each variance and covariance of observed variables is expressed as a function of model parameters. Model parameters are unknown and should be estimated. When a model is identified, each freely estimated parameter has a unique solution. This requires the number of unique elements in the observed covariance matrix ($p$), computed as $v \times (v+1)/2$ where $v$ is the number of observed variables, to be greater than or equal to the number of freely estimated parameters ($q$). The model degrees of freedom ($df$) equals $p$ minus $q$. If the requirement of $df \geq 0$ is not met, at least one model parameter does not have a unique solution (e.g., Kenny & Milan, 2012; MacCallum et al., 1993; Rindskopf, 1984). Mathematical determination of model identification is difficult and tedious. Some non-technical rules (e.g., Lee & Hershberger, 1990; MacCallum et al., 1993) have been proposed to help determine model identification for some SEMs. For any type of SEM, there are two basic requirements: 1) $df \geq 0$, and 2) each latent variable is assigned a metric. This is needed because each latent variable is unobserved and does not automatically have a metric. For factor analysis, at least three indicators are needed when there is only one factor, and at least two indicators are needed for each factor when there is more than one factor and factors are correlated. For models involving both measurement and structural parts and involving both covariance and mean structures, identification should be established for each part and each structure.

The structural regression model in our running example can involve the covariance structure only. Latent variables in this model include latent factors, disturbance, and measurement errors. Although not shown in Figure 32.1, the factor loading of an indicator of each factor, the coefficient from disturbance to each endogenous factor, and the coefficient from measurement error to each indicator are fixed to a constant (typically one) so that each latent variable is assigned a metric. This method of assigning a metric to latent variables is called unit loading identification. Other methods are available (e.g., unit variance identification; see Kline, 2016). The hypothesized model has $df = 561 - 70 = 491$. Specifically, the 33 measurement indicators provide $33 \times 34/2 = 561$ unique elements in the covariance matrix. The number of freely estimated model parameters is 70, which is the sum of the following categories:

- Variances of exogenous factors: 2
- Covariances of exogenous factors: 1
- Variances of disturbance associated with endogenous factors: 2
- Covariances of disturbance associated with endogenous factors: 0
- Variances of measurement errors: 33
- Covariances of measurement errors: 0
- Factor loadings: 29
- Path coefficients connecting latent factors: 3

The hypothesized structural regression model is theoretically identified. A model being theoretically identified does not guarantee that each parameter will have a unique solution when fitting to sample data. Empirical under-identification may be caused by unusual characteristics of data, for example a variable has a perfect linear relationship with the other variable. See Kenny and Milan (2012) for a non-technical discussion of model identification issues. Researchers should take a proactive approach to avoid empirical under-identification prior to data collection, for example by choosing indicators that have been carefully scrutinized for measuring latent constructs. When the psychometric properties of indicators are not well established in previous studies, it is suggested to include more than three indicators per factor.

Once the model is identified, the next step is to collect data from a representative sample and to estimate model parameters. For a covariance structure model, each variance and covariance of observed variables is expressed as a function of model parameters. If the model specification is correct, the covariance matrix of observed variables is identical to the one implied by the model in the
population. The discrepancy between these two matrices is quantified by fit function, denoted as $F$. Parameters are estimated through fitting the model to the sample data. For this step, it is crucial to choose an estimation method that is appropriate to the data at hand. Fit function is defined in different ways for different estimation methods. For continuous variables, the most commonly used method is maximum likelihood (ML). ML estimation adopts an iterative procedure to search for parameter estimates such that the likelihood of generating the data is maximized and, consequently, the fit function is minimized. This estimation method requires data following multivariate normal distribution. Other estimation methods are available for non-normally distributed continuous data or ordinal data, which are discussed later in the chapter. In this step, it is also important to determine the sample size needed for the planned analysis. SEM is a large sample size technique. Results from SEM analysis based on a small sample size can be misleading. For example, parameter estimates tend to be unstable, the test statistic may not have sufficient power to detect a misspecified model, and the chosen model is less likely to be replicated by other studies. Researchers have suggested that the ratio of sample size to the number of freely estimated parameters should not be smaller than ten (cf. Kline, 2016). However, such a suggestion should not be universally applied because sample size requirements for an SEM analysis depend on many factors, for example the number of variables in the model, model complexity, the characteristics of the data, and the strength of the relationship among variables. One suggested best practice is to conduct a power analysis to determine the sample size for a planned SEM analysis (MacCallum et al., 1996).

The next step is model evaluation and model comparison. When fitting a model to sample data, a test statistic $T$ is obtained: $T = (n - 1) \cdot \hat{F}$, where $n$ is sample size and $\hat{F}$ is the minimization of fit function. When the model is correctly specified and data meet assumptions corresponding to the chosen estimation method, the $T$ statistic follows asymmetrically a central chi-square ($\chi^2$) distribution with mean equal to model $df$ and variance equal to $2df$, and thus is often directly referred to as $\chi^2$ test statistic. This statistic is labeled as “chi-square” in Mplus and EQS, as “test statistic” in lavaan in R, and as “CMIN” in AMOS. If the $\chi^2$ test statistic from the sample analysis does not exceed the critical value of $\chi^2$ associated with the predefined alpha level, then the null hypothesis, that the covariance matrix is identical to the model-implied covariance matrix in the population, fails to be rejected. Otherwise, the null hypothesis is rejected. In addition to the $\chi^2$ test statistic, researchers often use some fit indices to assist model-data fit evaluation (these will be discussed later). If researchers have multiple competing models under consideration, $\chi^2$ difference test statistics may be used when the competing models are nested and information criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) may be used when the competing models are not nested.

When a researcher judges that the model-data fit is not satisfactory, he/she may opt to revise the model to improve model-data fit, for example based on modification indices. Modification indices give the approximate amount of reduction in $\chi^2$ test statistic that would result if a parameter is added to the model. Similar to the previous step, model re-specification based on a small sample size should proceed cautiously because the newly added parameters may simply be due to capitalization on chance characteristics of the sample data and may not render a meaningful interpretation. The recommendation in these instances is that model re-specification should be taken in a careful fashion and should not be guided by purely pursuing a good model-data fit.

Upon choosing a final model, the next step is to report and interpret parameter estimates. For the model in the running example, the parameters of interest are the three path coefficients connecting latent factors ($\hat{\beta}_1$ to $\hat{\beta}_3$). The point estimates of these parameters, along with their standard errors (SEs) or test statistics, can be reported. It is also informative to report $R^2$ for endogenous factors to indicate the predictive power of explanatory variables. In addition, the indirect effects from shame and guilt to achievement via motivation can be obtained. Indirect effects are computed as the product of path coefficients. For example, the estimate of the indirect effect from shame to achievement via motivation is computed as $\hat{\beta}_1 \cdot \hat{\beta}_3$ where $\hat{\beta}_1$ and $\hat{\beta}_3$ are the estimates of $\beta_1$ and $\beta_3$. A
well-known method for approximating SEs of an indirect effect is the delta method (Sobel, 1982). This method results in a z-statistic for testing $H_0: \beta_1 \ast \beta_3 = 0$. The delta method assumes that the sampling distribution of an indirect effect is normal, which is not realistic. Alternatively, bootstrapping methods (Efron, 1987) can be applied to construct an empirical sampling distribution of indirect effects. Various bootstrapping methods are available (cf. DiCiccio & Efron, 1996). In general, bootstrapping methods have been shown to outperform the delta method, particularly when the sample size is small.

**Factor Analysis: Confirmatory or Exploratory?**

The structural regression model formulated in our example includes both structural and measurement parts. Anderson and Gerbing (1988) suggested a two-step procedure and Mulaik and Millsap (2000) proposed a four-step procedure for assessing the fit of a structural regression model. Both procedures begin with factor analysis for evaluating a measurement model. The fit of the measurement model is compared to the fit of the structural regression model. Any deterioration in fit from the measurement model to the structural regression model is attributable to misspecification in the structural part. The main consideration of these procedures is that before testing directional relationships among latent factors, it is necessary to establish validity evidence for measurement instruments. For example, we need to ensure $V_1$–$V_{10}$ actually measure shame and $V_{11}$–$V_{20}$ measure guilt. Without this evidence, the estimated structural relationship (e.g., $\beta_1$ to $\beta_3$) may be misleading. Validity evidence should be established from different aspects. Relevant to SEM is the use of factor analysis to provide supporting evidence of construct validity. For example, $V_1$–$V_{10}$ should have sufficiently large loadings on shame but not on guilt, and the correlation coefficient between shame and guilt should not be too high. Instead of detailing procedures for assessing structural regression models, below we comment on the use of factor analysis for two reasons. First, it is the first step for assessing a structural regression model (and other SEMs involving latent factors measured by measurement indicators). Second, it is a frequently used statistical procedure to study the psychometric properties of an instrument.

When conducting factor analysis, researchers may choose to use exploratory factor analysis (EFA; Spearman, 1904) or confirmatory factor analysis (CFA; Jöreskog, 1969). EFA is a data-driven technique and it is appropriate when no strong assumptions about factor structure are available. EFA, often accompanied by an appropriate rotation method, is particularly useful during the process of scale development and scale refinement. Because of the data-driven nature, the factor structure obtained from EFA needs to be cross-validated through either EFA or CFA. If the sample size is large, researchers may split the sample into two independent subsamples, one for studying the factor structure and the other for validation. Alternatively, cross-validation can be achieved by collecting a new independent sample. Different from EFA, CFA is a theory-driven technique requiring users to specify the number of factors and the pattern of factor loadings (i.e., fixed vs. freely estimated). CFA is useful when measurement indicators have been carefully scrutinized in other studies.

Researchers may prefer CFA to EFA for various reasons. From a measurement point of view, it is preferable that each measurement indicator loads on only one latent factor. A factor model with such a simple structure is referred to as an independent cluster model (ICM; Jöreskog, 1969). Figure 32.2 demonstrates a key distinction between an EFA model and an ICM–CFA model using shame and guilt as an example. The primary loadings are indicated by solid lines while cross-loadings are indicated by dashed lines. An EFA model includes both primary loadings and cross-loadings while an ICM–CFA model removes all cross-loadings. If all cross-loadings are truly zero or close to zero, the factor structure obtained from EFA can be validated by an ICM–CFA model. However, not all measurement instruments conform to an ICM structure; thus, fitting an ICM–CFA model often results in model-data misfit. In these circumstances, researchers often revise the initial ICM–
CFA model by removing indicators that cross-load on untargeted factors from analyses, fixing cross-loadings less than a certain cut-off point (e.g., 0.3) to be zero (Cudeck & O’Dell, 1994), and sometimes changing the number of underlying factors. All these approaches could be problematic. First, removing indicators that cross-load on untargeted factors may result in construct under-representation. The construct defined by the remaining indicators can be much narrower than the original one and, consequently, the tested relationships among latent factors can be misleading. Therefore, this approach should not be taken without evidence from item content analysis. Second, fixing cross-loadings that are less than a certain cut-off point to zero may lead to model under-specification. Consequently, the model-data fit is deteriorated, factor correlations are inflated, and path coefficients connecting factors are biased (e.g., Asparouhov & Muthén, 2009). Researchers may subsequently execute extensive model modifications including opportunistically adding new factors or changing factor loading patterns (e.g., make \( V \) as a measure of guilt) to find a well-fitting model. Such actions lead to an accumulation of specification search errors, particularly when the sample size is not sufficiently large (MacCallum et al., 1992).

The tradeoff between model parsimony and better-presenting substantive theories has been an issue until recently when EFA was integrated into the SEM framework, termed “exploratory structural equation modeling” (ESEM; Asparouhov & Muthén, 2009) and implemented in Mplus. As such, EFA can take full advantage of the developments in SEM such as obtaining model fit indices, conducting statistical testing for rotated factor loadings, and being tested for measurement invariance across groups. At the same time, it retains flexibility by allowing the existence of many non-ignorable cross-loadings. ESEM may be a choice when an instrument measures a multidimensional construct but has a complex factor loading matrix (i.e., one containing many nonzero cross-loadings).

**Choice of Estimation Methods**

Once data are collected, researchers need to choose an estimation method appropriate for the data at hand. In ID research, response data are often non-normal or ordered categorical with a limited number of response options. Applying an ML estimation method to non-normal or ordinal data may lead to biased parameter estimates, SEs, model \( \chi^2 \) test statistics, and fit indices. The degree of bias depends on how far continuous variables deviate from normality (e.g., Satorra & Bentler, 1994), the number of response categories, and the distribution of ordinal variables. Applying an ML method to ordinal variables is less problematic as the number of response points increases and the distribution of ordinal variables approaches symmetricity (e.g., Rhemtulla et al., 2012, Yang & Green, 2015).
For non-normal continuous data, two approaches may be considered. One approach is to use a weighted least squares (WLS, aka ADF; Browne, 1984; Muthén, 1984) estimation method which does not rely on the normality assumption. Previous studies (e.g., Curran et al., 1996) have shown that WLS performed reasonably well only when the sample size was large (e.g., >1,000), which is impractical for many small-scale research studies. Another approach is to use robust maximum likelihood methods (Asparouhov & Muthén, 2010; Satorra & Bentler, 1994; Yuan & Bentler, 2000). This approach gives identical parameter estimates to those from ML, but corrects for SEs of parameter estimates, \( \chi^2 \) test statistics, and \( \chi^2 \)-based-fit indices. Various correction methods have been proposed, which differ in how to estimate robust SEs, how to adjust the \( \chi^2 \) test statistic, and whether to allow missing data. These correction methods are implemented in Mplus under different estimators labeled as MLF, MLM, MLMV, and MLR (cf. Maydeu-Olivares, 2017; Savalei, 2010). Maydeu-Olivares (2017) showed that MLMV outperformed the other methods.

For ordered categorical data, particularly when the number of response options is small, researchers have suggested fitting models to polychoric correlations and thresholds. Polychoric correlations are correlations among continuous variables that underlie observed ordinal variables obtained via applying cut-off values (called thresholds) to continuous variables. Muthén et al. (1997) proposed fitting models to polychoric correlations and thresholds using a diagonally weighted least squares (DWLS) and an unweighted least squares (ULS) estimation methods, both using a fit function that is simpler than those in WLS. This simplification leads to model \( \chi^2 \) test statistics deviating from a central \( \chi^2 \) distribution when the model is correctly specified. Therefore, robust corrections are needed (Asparouhov & Muthén, 2010; Muthén, et al., 1997). In both Mplus and the lavaan package in R, DWLS is implemented as WLSM and WLSMV, and ULS is implemented as ULSMV. Studies found that parameter estimates, SEs, and \( \chi^2 \) test statistics from ULS tended to be more accurate compared to DWLS (e.g., Savalei & Rhemtulla, 2013).

**Use of Fit Indices for Model-Data Fit Evaluation**

Numerous fit indices have been proposed to assist model-data fit evaluation. The most commonly used global fit indices include the root mean square error of approximation (RMSEA; Steiger, 1990; Steiger & Lind, 1980), comparative fit index (CFI; Bentler, 1990), Tucker-Lewis index (TLI, aka non-normed fit index, NNFI; Tucker & Lewis, 1973; Bentler & Bonett, 1980), and standardized root mean square residual (SRMR; Bentler, 1995). RMSEA, CFI, and TLI are \( \chi^2 \)-based-fit indices. RMSEA measures how far the fit of the hypothesized model is from that of a perfectly fitting model, and incorporates a penalty term for making a model too complex. A smaller value indicates a better model-data fit. CFI and TLI are often referred to as incremental fit indices. They measure the degree to which the hypothesized model improves fit over the baseline model. A baseline model assumes no relationship among variables, which often leads to the worst model-data fit. The population CFI and TLI values are in the range of 0 and 1. A greater value indicates a better model-data fit. SRMR takes a value from 0 to 1. A smaller SRMR value indicates a better model-data fit.

Applied researchers often adopt some cut-offs of fit indices (e.g., RMSEA < 0.06, CFI > 0.95, and TLI > 0.95; Hu & Bentler, 1999) to evaluate model-data fit. Studies have repeatedly shown that there is no single set of cut-offs applicable to all models under all conditions (e.g., Garrido et al., 2016; Marsh et al., 2004; Yang & Xia, 2015), particularly when data are ordinal (e.g., Brosseau-Liard & Savalei, 2014). Using both analytical approaches and computer-generated data, Xia and Yang (2018, 2019) showed that the behavior of fit indices under DWLS and ULS depends not only on the level of model misspecification but also on the number of response categories and threshold values. Their studies also showed that different fit indices can be obtained by fitting the same model to the same dataset with different estimation methods. It is, therefore, suggested that users of SEM should not try to universally apply a specific set of cut-offs of fit indices and pursue solely goodness-of-fit. Instead, researchers should use a holistic approach by weighing in on the decision.
with multiple criteria including model $\chi^2$ test statistic, fit indices, modification indices, inspecting the residual covariance matrix, and expected parameter changes. More importantly, model specification and re-specification should be guided by theory (e.g., Garrido et al., 2016; Kline, 2016; Xia & Yang, 2018).

**Use of Composite Variables Versus Measurement Indicators**

In the model specification step, it is important for a researcher to decide how each latent construct is represented in the hypothesized model. This decision impacts the determination of model degrees of freedom, sample size, and choice of estimation method. The structural regression model depicted in Figure 32.1 is complex with $df = 491$. Often seen in SEM applications is that a composite variable, for example by taking sum or mean scores across measurement indicators, is used as an approximation of the latent construct. The resulting model is a path analytic model, as depicted in Figure 32.3.

Equations corresponding to this path model are:

\[
C_3 = \beta_{0i} + \beta_1 C_1 + \beta_2 C_2 + D_i
\]

\[
C_4 = \beta_{0i} + \beta_3 C_3 + D_i,
\]

where symbols with superscript C indicate those for composite variables. Equation 32.3 is in the same form as Equation 32.1 except that the composite variables $C_1$–$C_4$ replaced $F_1$–$F_4$. Compared to the model with all measurement indicators explicitly included, this approach dramatically reduces model complexity with $df = 2$. However, it makes a very restrictive assumption: composite variables perfectly represent latent variables, that is, the reliability coefficient of composite scores is one. When this condition is not met, which is typical in practice, path coefficients based on the path analysis (i.e., $\beta_1$ to $\beta_3$) differ from those based on the structural regression model (i.e., $\beta_1^c$ to $\beta_3^c$). The direction of bias is hard to predict, depending on the reliability coefficient of composite scores, the complexity of the model, and whether the model is correctly specified or misspecified (Cole & Preacher, 2014).

For studies based on a small sample size relative to model size, reducing model size and using composite scores as an approximation of latent factors is tempting. In these circumstances, to correct for bias due to measurement errors, researchers may use a structural regression model with a composite variable as the single indicator for each factor, and constrain the proportion of measurement error variance for the indicator as one minus the reliability coefficient. Figure 32.4 shows the application of this approach to the running example. In the longitudinal mediation modeling framework, Zhang...
SEM and Factor Analysis

and Yang (2020) showed analytically that this approach yielded unbiased path coefficients connecting factors (i.e., $\beta_1$ to $\beta_3$). It should be noted that this approach requires using an appropriate method to estimate the reliability coefficient of composite scores. If the scale conforms to a one-factor model, then coefficient $\omega$ is appropriate. If the scale measures one major factor but also contains minor factors, researchers may fit the measurement indicators to a bifactor model and compute coefficient omega hierarchical ($\omega_H$). If the scale measures multiple subdomains of a construct, then computing a single composite variable across all indicators can be problematic. In such circumstances, researchers may consider item-parceling by creating parcels across a subset of indicators and using these parcels as indicators of factors (cf. Bagozzi & Edwards, 1998; Little et al., 2002).

Contributions to ID Research

SEM is a flexible statistical framework allowing researchers to investigate complex interrelationships among variables. It enables ID researchers to investigate the influence of individual and contextual factors on learning outcomes, to study the psychometric properties of measurement instruments, to compare group differences in learning outcomes, to examine learning trajectories over time, and to study how a learning trajectory is related to individual and contextual factors, to name a few applications. Several examples of SEM applications in ID research are as follows. Teimouri (2018) conducted exploratory factor analyses to examine the factor structure of a newly developed scenario-based questionnaire for measuring shame and guilt. The developed questionnaire was then used in a second study to investigate the effects of shame and guilt on L2 learners’ motivation and language achievement using multiple regression. Lee and Evans (2019) employed path analysis to examine the mediating role of writing self-regulatory efficacy and apprehension in the relationship between perceived usefulness of giving peer feedback and L2 writing self-efficacy. Bang and Hiver (2016) used both confirmatory factor analysis and structural regression models to examine the relationships among L2 linguistic knowledge, L2 listening strategy use, self-determined motivation, and L2 listening anxiety relative to L2 listening proficiency, each of which was measured by multiple measurement indicators. Noels et al. (2019) applied latent growth curve modeling and cross-lagged panel

Figure 32.4 Structural Regression Model with Single Indicator Corrected for Measurement Errors. Note: Var($C_1$) to Var($C_4$) indicate variance of composite scores, and $\rho_{C1}$ to $\rho_{C4}$ indicate reliability of composite scores.
model to longitudinal data collected in a language course to study the relationship between self-determined motivation and engagement over time.

Other applications of SEM in ID research are emerging. Ardasheva et al. (2019) conducted a review of articles published in *Studies in Second Language Acquisition*, *Language Learning*, and the *Modern Language Journal* from January 2007 to September 2017. These three journals were chosen because they publish quantitative studies on IDs. The authors found that a total of 46 articles applied SEM. Some articles involved more than one study and/or applied more than one type of SEM, resulting in a total of 51 studies and 161 models. About 72% of these studies applied EFA/CFA to examine the factor structures of measurement tools or to establish the reliability and validity of measurement tools. The most frequently studied variables are motivation, cognition, linguistic knowledge, social perceptions, grammar, vocabulary, metalinguistic knowledge, metacognition, self-regulation, and attitudes. The majority of these studies (70%) focused on adult populations and only a few were specific for K-12 populations. We conducted a brief further review of articles published in the year 2019 in *Studies in Second Language Acquisition*. The overall pattern is similar. We found that only seven studies applied SEM and five of them used EFA/CFA to study the psychometric properties of measurement tools (e.g., measuring linguistic knowledge and L2 motivation). Only one study included participants younger than 18 years old. Although not as frequent as EFA/CFA, other types of SEM have also been applied in ID research, including path analysis, structural regression models, longitudinal models, multiple-group SEM, etc. Many studies employed more than one type of SEM, for example by using factor analysis to establish the construct validity of measurements and then applying path analysis or structural regression modeling to investigate the relationship among constructs.

**Future Directions**

SEM provides great potential for ID researchers looking to uncover the interrelationships among variables in ways that are not easily achievable using classical statistical procedures, such as regression analysis. We expect that the applications of SEM in ID research will continue to increase. Note that the benefits of using SEM are built upon the assumption that SEM is properly used. As we laid out in the technical features section, a typical SEM analysis involves five steps, from model specification to results reporting, in which various decisions ought to be made, such as sample size determination, choice of estimation method, level of analysis, and model-data fit evaluation. Unfortunately, there is no golden rule for any of these inquiries. A proper decision requires users to have a solid understanding of various aspects of SEM analyses, including being sufficiently knowledgeable on measurement theory. It is not uncommon that SEM is misused or not properly conducted in the current ID research. For example, ordinal measurement indicators with a limited number of response categories (i.e., three-point Likert-type scale) are treated as continuous. We have discussed alternative methods for ordinal data in the chapter. It is also not uncommon for SEM analyses to be based on small sample sizes. The review of journal articles by Ardasheva et al. (2019) reveals that 24% and 49% of SEM studies were based on less than 100 and 200 participants, respectively. One article we reviewed in 2019 in *Studies in Second Language Acquisition* used a sample size of around 150 to examine complex longitudinal models and conducted measurement invariance tests. This is not optimal, and as we discussed previously in the chapter, such analyses may result in low statistical power to detect misspecified models, as such interpretations can be misleading and impact theory testing. Furthermore, studies in IDs may involve student populations, where students are nested within classrooms and schools, thus creating clustering effects. However, none of the studies reviewed in Ardasheva et al. (2019) explicitly discussed and handled clustering effects. ID researchers are encouraged to take various aspects of analyses into consideration when planning an SEM study. Readers may also consult with articles targeting improved SEM practices for applied researchers (e.g., Hancock & Schoonen, 2015; Ockey & Choi, 2015; Schreiber et al., 2006).
We end this chapter by providing two additional SEM techniques that ID researchers may consider in future research. Both techniques may alleviate some concerns due to clustering effects or small sample sizes. One is multilevel SEM (cf. Stapleton, 2013). This technique enables researchers to differentiate variability attributable to clustering effects (e.g., students nested within classes) and random sampling error when studying relationships among variables. The overall procedure for conducting multilevel SEM is similar to what we reviewed earlier in the chapter. However, model specification and model evaluation should be done for each level (e.g., student level and classroom level). The second technique is Bayesian SEM (Muthén & Asparouhov, 2012). Different from conventional SEM which treats model parameters as fixed, Bayesian SEM applies Bayes’ theorem to SEM, treats model parameters as random variables, and incorporates researchers’ prior knowledge into modeling. One important yet challenging task for conducting Bayesian SEM is the specification of prior distribution for each parameter. Prior distribution can be informative or noninformative, depending on the researcher’s belief or knowledge of the relationship between variables. When used properly, Bayesian SEM can handle complex model specifications and has advantages over the conventional SEM in studies based on small sample sizes.

Notes

1 For some constructs, the relationship between measurement indicators and latent variables may be opposite with indicators being the cause of latent variables. Such type of measurement indicator has been referred to as formative indicators (cf. Bagozzi, 2007; Howell, Breivik, & Wilcox, 2007). Bollen and Bauldry (2011) further classified formative indicators into three types: causal indicators, composite indicators, and covariates.

2 The example is based on the research scenario presented in Teimouri (2018) with some modifications.

References

Arbuckle, J. L. (2019). Amos (version 26.0) [computer program]. Chicago, IL: IBM SPSS.


