4

INTRODUCTION TO STRUCTURAL EQUATION MODELLING

Chapter Overview

Structural equation modelling (SEM) refers to a family of techniques, including (but not limited to) path analysis, **confirmatory factor analysis (CFA)**, structural regression models, autoregressive models and latent change models (Marcoulides & Schumacker, 2001; Raykov & Marcoulides, 2000). SEM utilises the analysis of covariances to examine the complex interrelationships within a system of variables. SEM builds upon, integrates and greatly expands the capabilities of more traditional statistical techniques such as multiple linear regression and ANOVA (Hoyle, 2012).

This chapter provides a conceptual introduction to SEM. We describe the advantages of SEM, outline the assumptions and requirements for SEM, define key terms and concepts and provide brief, non-technical introductions to path analysis and **latent variables**. In Chapter 5, we delve into more detail, and we discuss specification, **identification** and estimation in SEM. In addition, we discuss Wright's rules and systems of equations. In Chapter 6, we discuss model building and provide an example where we fit and estimate a **hybrid model**.

Advantages of SEM

SEM is extremely versatile; it places very few restrictions on the kinds of models that can be tested (Hoyle, 1995, 2012). Consequently, SEM allows researchers to test a wider variety of hypotheses than would be possible with most traditional statistical techniques (Kline, 2015). Using SEM to specify a given model based on theory, researchers can examine the degree to which the model can reproduce the relationships among **observed variables** (and patterns of means). Alternatively, researchers can test competing theories by fitting alternative models to determine which of the competing models appears to best fit the observed data (Kline, 2015).

SEM allows researchers to distinguish between observed and *latent* variables and to explicitly model both types of variables. Latent variable models have both conceptual and statistical advantages over traditional observed variable techniques. Using latent variables, researchers can include **latent constructs** in their analyses. A **construct** is a concept, model or schematic idea (Bollen, 2002). Thus, *latent constructs* are non-observable concepts, such as cognitive ability, self-concept or optimism. Although latent constructs themselves are not directly observed, their presence and influence can be inferred based on variables that are directly observed. For example, educators use observable indicators such as test scores, self-reports, teacher ratings and/or behavioural observations to infer latent constructs such as students' academic engagement. Latent variable models permit a level of methodological and theoretical freedom that is nearly impossible in most other statistical analyses.

Furthermore, using latent variables in SEM accounts for potential errors of measurement, allowing researchers to explicitly account for (and model) **measurement error**

(Raykov & Marcoulides, 2000). The ability to separate measurement error or 'error variance' from 'true variance' is one of the reasons that SEM provides such powerful analyses. In multiple regression, measurement error within a predictor variable attenuates the regression weight from the predictor variable to the dependent variable, downwardly biasing the parameter estimates (Baron & Kenny, 1986; Campbell & Kenny, 1999; Cole & Preacher, 2014). Structural equation models that include latent variables use multiple indicators to estimate the effects of latent variables. This approach corrects for the unreliability within the measured predictor variables, providing more accurate estimates of the effects of the predictor on the criterion.

Measurement error in the **mediator** (or other variables in the mediational model) can also produce biased estimates of direct, indirect and **total effects** (Baron & Kenny, 1986; Cole & Preacher, 2014). Cole and Preacher (2014) outline four serious consequences of ignoring measurement error in mediational analyses: (1) measurement error can cause path coefficients to be either over- or underestimated (and predicting the direction of bias becomes quite difficult as the complexity of the model increases), (2) measurement error can decrease the power to detect incorrect models, (3) even seemingly small amounts of measurement error can make valid models appear invalid and (4) differential measurement error across the model can actually change substantive conclusions. Generally, these four issues become increasingly problematic as model complexity increases (p. 300). Fortunately, using latent variables as the structural variables in mediational models eliminates these issues: using latent variables in SEM accounts for the measurement error and produces unbiased estimates of the direct, indirect and total effects.

Finally, SEM allows researchers to specify a priori models and to assess the degree to which the data fits the specified model. SEM provides a comprehensive statistical approach to test existing hypotheses about relations among observed and latent variables (Hoyle, 1995). In this way, SEM forces the researcher to think critically about the relationships among the variables of interest and the hypotheses being tested. Further, SEM allows researchers to test competing theoretical models to determine which model best reproduces the observed variance–covariance matrix.

SEM analyses can include means as well as covariances. In fact, researchers can also use SEM techniques to model *latent* means. Thus, SEM provides a framework for examining between-group differences: **multiple group SEM (MG-SEM)** enables between-group comparisons of any model parameters, including latent means. Therefore, MG-SEM facilitates the examination of both differences in patterns of interrelationships among variables across groups and differences in means and variances across groups. Latent growth curve models also incorporate means into the structural equation model (Bollen & Curran, 2004; Duncan et al., 1999; Grimm et al., 2016). For the remainder of this section (Chapters 4–6), we confine our discussion to modelling covariance structures without means. Chapter 7 introduces means structure analysis in the context of latent growth curve models to study change across time.

Assumptions and requirements of SEM

SEM is a regression-based technique. As such, it rests upon four key assumptions and requirements necessary for any regression-based analyses: normality, linearity, independence and adequate variability. We briefly discuss each of these requirements below.

Normality. Many of the assumptions of SEM parallel those of multiple linear regression. Namely, SEM assumes that the variables of interest are drawn from a multivariate normal population (Hancock & Mueller, 2006; Kaplan, 2000). ML estimation performs optimally when the data is continuous and normally distributed (Finney & DiStefano, 2006; Kaplan, 2000). Generally, SEM is fairly robust to small violations of the normality assumption; however, extreme non-normality can cause problems. (For more information about dealing with non-normal data, see Curran et al., 1996; Finney & DiStefano, 2006; Hayakawa, 2019; S. G. West et al., 1995.)

Linearity. As in multiple regression, SEM assumes that the variables of interest are linearly related to each other. In addition, we can examine non-linear and interaction effects using SEM. Interested readers should consult Maslowsky et al. (2015) for a comprehensible introduction to modelling interaction effects in SEM.

Sampling. ML estimation assumes that the data represents a simple random sample from the population. Chapter 1 of this book discussed the pitfalls of assuming that data are independent. Multilevel modelling provides a solution to issue of non-independence. Multilevel SEM combines multilevel modelling and SEM techniques to analyse data that have been collected using multistage or cluster sampling techniques (Kaplan, 2000). However, we need to walk before we run, so our introduction to SEM assumes that data are independent. Space precludes us from discussing multilevel SEM in this book. For more information about multilevel SEM, we recommend Bauer et al. (2006), Heck and Thomas (2000, 2015), Hox et al. (2017), Muthén and Muthén (1998–2017), Preacher, Zhang et al. (2011, 2016) and Preacher, Zyphur et al. (2010).

Sample Size. Because standard SEM uses ML estimation to minimise the discrepancy between the observed covariance matrix and the model-implied covariance matrix, SEM is a large sample technique. (For information about alternative estimation methods, see Kaplan, 2000, or Hoyle, 1995.) There are no definitive rules for a minimum sample size, but the literature provides general rules of thumb. Under most circumstances, sample sizes below 100 are too small to use SEM techniques (Kline, 2015; Schumacker & Lomax, 1996). Latent variable models generally require larger sample sizes than comparable path models (that include only observed variables). In general, models with larger numbers of freely estimated parameters require larger sample sizes. As the ratio of the number of cases to the number of parameters declines, the estimates generated by SEM become more unstable (Kline, 2015). Kline (2015) recommends maintaining a ratio of at least 10 cases per freely estimated parameter. In many instances, sample sizes of 200 or more are sufficient for estimating structural equation models, especially if the variables are normally distributed and obtained from a random sample of subjects. Sample sizes of 250 to 500 are typical in SEM studies (Kline, 2015). In summary, SEM generally requires large samples; however, the minimum sample size ultimately depends on the size and complexity of the structural equation model.

Range of values. Because SEM is essentially a correlational technique, anything that affects the magnitudes of the covariances among the variables in the model impacts the SEM analysis. For example, restriction of range in one variable attenuates the covariance between that variable and any other variables in the model. Similarly, a small number of influential outliers can have a large impact on the variance–covariance matrix. Such issues are likely to lead to biased parameter estimates.

Understanding SEM

The basic building block of any structural equation model is the variance–covariance matrix. (Means are also required for some analyses such as growth curve analysis and MG-SEM, but to lay the groundwork of SEM, we begin with traditional covariance structure analysis.) SEM provides a way to use the covariance matrix to explain complex patterns of interrelationships among variables. In fact, the covariance matrix is the **sufficient statistic** for standard structural equation model: it contains all the information needed to fit a standard SEM analysis. We can compute the covariance matrix using the correlation matrix and the standard deviations for each variable. Because it is possible to create and analyse a structural equation model without the raw data file, interested researchers can conduct SEM using a published covariance or correlation matrix.[1] The covariance matrix is the unstandardised version of the correlation matrix; therefore, it is simple to create the covariance matrix using standard deviations and correlations. As we delve into what might seem to be a complex barrage of symbols, jargon and numbers, remember that the covariance matrix is at the heart of SEM.

SEM's versatility stems from the fact that it incorporates path analysis (or simultaneous equation modelling) and factor analysis (or latent variable modelling) into one modelling framework. We begin our tour of SEM by providing a brief definition and overview of path analysis and factor analysis. In the chapters that follow, we demonstrate how path analysis and factor analysis can be combined into a single latent variable modelling framework.

[1] Technically, it is considered proper form to analyse a covariance matrix, but under a variety of conditions, analysing a correlation matrix will produce the same results, as a correlation matrix is simply a standardised version of a covariance matrix.

What is path analysis?

More than 100 years ago, Sewall Wright (1920, 1923) developed path analysis, a technique that allows for the estimation of a system of simultaneous equations (Bollen, 1989). In traditional statistical analyses such as regression, a variable can serve as either an independent variable or a dependent variable; however, the same variable cannot serve as both a predictor and an outcome simultaneously (Hoyle, 2012). Mediators are a classic example of variables that serve as both independent and dependent variables, and mediational models cannot be estimated in a single multiple regression analysis. In contrast, in path analysis, a single variable can act as both an independent variable and a dependent variable. Therefore, path models can specify a complex system of predictive pathways among a large number of variables.

Path diagrams

Path diagrams, visual displays of structural equations, are perhaps the most intuitive way to conceptualise the process of developing and testing a specified model. Most predictive models can be represented as path models. Exhibit 4.1 provides a summary of the typical symbols in a path diagram.

Exhibit 4.1 Symbols in a path diagram

Observed variables	□	Paths (direct effects)	→
Latent variables	○	Covariances or correlations	↔

Note. The double-headed arrow represents a covariance in the unstandardised solution and a correlation in the standardised solution.

An observed (or manifest) variable is a variable that is actually measured. For example, a student's score on a test or a subscale is an observed variable. In a path diagram, rectangles indicate observed or measured variables. In contrast, circles or ellipses represent latent variables, which are not directly observed in the sample (data). Straight single-headed arrows represent paths. Just as in multiple regression, these paths represent the degree to which the predictor variable predicts a given outcome variable *after controlling for* (or holding constant) the other variables that also contain direct paths to (arrows pointing to) the dependent variable. In fact, we can construct a path diagram to display any multiple regression model. Double-headed arrows, which are generally curved, indicate a simple bivariate correlation between two variables.

Figure 4.1 illustrates a simple three-variable mediation model as a path diagram. The three observed variables are growth mindset, academic persistence and academic achievement. Straight single-headed arrows, called paths, connect growth mindset → academic persistence, academic persistence → academic achievement and growth mindset → academic achievement. The direction of the arrowhead is important: the arrow points from the predictor variable to the outcome variable.

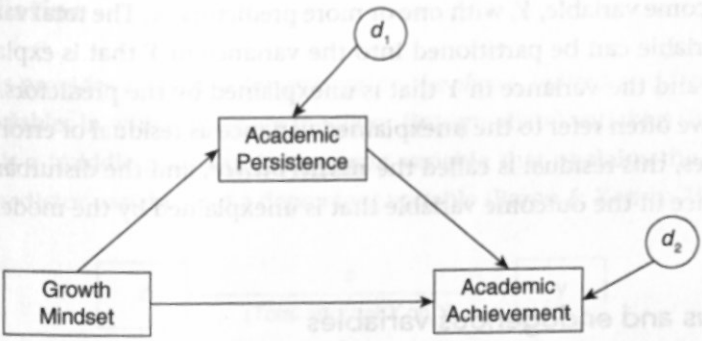

Figure 4.1 A path model in which growth mindset predicts academic persistence, which in turn predicts academic achievement

Just as in regression, path coefficients in SEM can be unstandardised or standardised. **Unstandardised path coefficients** depict the expected unit of change in the dependent variable given a one-unit change in the predictor variable, holding the other variables in the model constant. Unstandardised path coefficients or parameters reflect the scales of measurement of both the independent and dependent variables. Therefore, the interpretation of unstandardised path coefficients depends on the scales of both the predictor and criterion variables. In contrast, **standardised path coefficients** are analogous to beta coefficients in regression; conceptually they represent path coefficients in a model where all variables are standardised (i.e. *z* scores with mean = 0 and standard deviation/variance = 1).

Each of the parameters has an associated standard error, and we can test the statistical significance of every parameter that we estimate. As in multiple regression, each unstandardised path coefficient is divided by its standard error to compute a critical ratio. If the absolute value of this ratio is greater than or equal to 1.96, the path is statistically significant (at $\alpha = .05$). If the ratio of the unstandardised path coefficient to its standard error is less than 1.96, the path is considered non-statistically significant.

For example, the path from growth mindset → academic achievement indicates that growth mindset predicts academic achievement. Because both growth mindset and academic persistence predict academic achievement, the growth mindset → academic achievement path represents the direct *effect* of growth mindset → academic achievement, after controlling for academic persistence. Similarly, the academic persistence → academic achievement path represents the direct effect of academic persistence → academic achievement, after controlling for growth mindset. Only growth mindset predicts academic persistence; therefore, the path from growth mindset → academic persistence does not control for any other variables.

The latent variables, d_1 and d_2 in Figure 4.1 represent *disturbance variances*. We predict the outcome variable, *Y*, with one or more predictors, *X*. The total variance in the outcome variable can be partitioned into the variance in *Y* that is explained by the predictor(s) and the variance in *Y* that is unexplained by the predictors. In multiple regression, we often refer to the unexplained variance as residual or error variance. In path analyses, this residual is called the ***disturbance***, and the disturbance variance is the variance in the outcome variable that is unexplained by the model.

Exogenous and endogenous variables

SEM makes a key distinction between **exogenous variables** and **endogenous variables**. *Exogenous* variables predict other variables, but they are not predicted by any other variables in the model. In our simple example, growth mindset is an exogenous variable: it is purely a predictor variable. Exogenous variables may be (and generally are) correlated with any other exogenous variables, and they predict one or more variables in the model. However, we assume the causes of exogenous variables (or variables that explain the variance in the exogenous variables) lay outside the model.

In contrast, *endogenous* variables are predicted by one or more variables in the model. Just as in regression, every endogenous variable in the model contains a residual (called a disturbance), representing the unexplained variance in the variable. Therefore, the total variance in academic persistence equals the variance that is explained by growth mindset plus the disturbance (unexplained) variance. In path analysis, endogenous variables can also predict other endogenous variables. In other words, a variable can be a predictor only (exogenous), a predictor and an outcome (endogenous) or an outcome only (endogenous). In our simple model, both academic persistence (which is a mediator) and academic achievement (which is an outcome only) are endogenous variables.

Walking through our simple conceptual example in Figure 4.1 illustrates how the same variable can be both a predictor and an outcome. (This example assumes all variables are observed.) Academic persistence predicts subsequent academic achievement.

Students who are more persistent tend to have higher academic achievement. Academic persistence is the predictor and academic achievement is the outcome. However, growth mindset predicts academic persistence (Dweck et al., 2014). Here, growth mindset is the predictor and academic persistence is the outcome variable. Growth mindset predicts academic persistence, which in turn predicts academic achievement. Thus, academic persistence is both an outcome variable and a predictor: academic persistence is predicted by growth mindset and a predictor of academic achievement.

Estimating direct, indirect and total effects in path models with mediation

Path analysis provides a method for estimating the *direct*, *indirect* and *total effects* of a system of variables in which there are mediator (intermediate) variables (Bollen, 1989). A mediator is a 'middle man', an intervening variable that explains the relationship between a predictor variable and a dependent variable (Baron & Kenny, 1986).

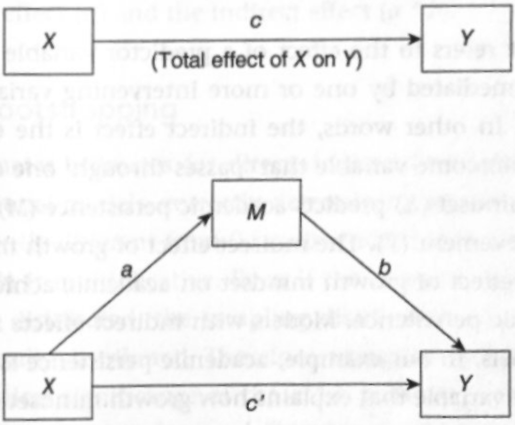

Figure 4.2 A simple mediational model

Direct effects

A **direct effect** represents the independent contribution of a predictor variable (*X*) on an outcome variable (*Y*), after controlling for all of the other variables that also predict *Y* (and share variance with *X*). In our simple example, growth mindset is *X*, academic persistence (which is both a predictor and an outcome) is *M* and academic achievement (which is only an outcome) is *Y*. In multiple regression, the **partial regression coefficient** is a direct effect: it is the effect of *X* on *Y*, after controlling for all the predictor variables in the model. In our model above, the direct effect of growth mindset on academic achievement is the effect of growth mindset on academic achievement, after controlling for academic persistence. If this direct effect is 0, then growth mindset does not predict academic achievement after controlling for the effects of academic persistence on academic achievement. In other words, any variance in academic achievement that is explained by growth mindset is also explained by academic persistence. Therefore, once we control for academic persistence, growth mindset does not predict any additional variance in academic achievement. If there is a direct effect of growth mindset on academic achievement, then growth mindset explains additional variance in academic achievement, over and above the amount that is explained by academic persistence. This direct effect is just like a partial regression coefficient in a multiple regression equation. In fact, if we ran a multiple regression with academic persistence and growth mindset as predictors of academic achievement, the partial regression coefficient for growth mindset would be identical to the direct effect from the path analysis.

Indirect effects

An **indirect effect** refers to the effect of a predictor variable (*X*) on an outcome variable (*Y*) that is mediated by one or more intervening variables (*M*) (Raykov & Marcoulides, 2000). In other words, the indirect effect is the effect of the predictor variable on the outcome variable that 'passes through' one or more intervening variables. Growth mindset (*X*) predicts academic persistence (*M*), which in turn predicts academic achievement (*Y*). The indirect effect of growth mindset on academic achievement is the effect of growth mindset on academic achievement that is also shared with academic persistence. Models with indirect effects are often referred to as mediational models. In our example, academic persistence is a *mediator* variable: it is an intermediate variable that explains how growth mindset influences academic achievement (Baron & Kenny, 1986). Figure 4.2 illustrates a simple mediational model with an indirect effect of *X* on *Y* via *M*. The coefficient for path from *X* to *M* is *a* and the coefficient for the path from *M* to *Y* is *b*. The product of the two paths (*a* * *b*) provides an estimate of the indirect effect.

Indirect effects do not exist in *standard* multiple regression models (in which variables are either predictors or outcomes, but not both), but they do exist in path analysis (and SEM). Recursive path models can be estimated using multiple regression analyses in a traditional OLS framework. However, using multiple regression to estimate the indirect effect requires estimating two separate regression models. The first model regresses *Y* on *X*. The second model regresses *Y* on *M* and *X*. The indirect effect of *X* on *Y* (via *M*) is the effect of *X* on *Y* (the total effect) – the effect of *X* on *Y* after controlling for *M* (the direct effect). Because the Total effect = Direct effect + Indirect effect, the indirect effect of *X* on *Y* via *M* is the difference in those two coefficients. When we refer to path analysis and SEM, we are referring to single-step methods for estimating these models. When we refer to 'standard multiple regression analysis', we are referring to the process of running a single multiple regression model.

Total effect

The total effect of a predictor variable (*X*) on an outcome variable (*Y*) is the effect of *X* on *Y*, whether or not it is mediated by a third variable, *M*. There are two ways to compute the total effect. The first is quite simple: the simple linear regression of *Y* on *X* produces the total effect of *Y* on *X*. The second method to compute the total effect is to sum the direct and indirect effects. In other words, the total effect of *X* on *Y* is the sum of the direct effect of *X* on *Y* and the indirect effect of *X* on *Y* that is mediated by the intermediate variable, *M*. In the top panel of Figure 4.2, the total effect is *c*, the regression coefficient for the model that regresses *Y* on *X* but does not include *M*. Alternatively, in the bottom panel of Figure 4.2, we can compute the total effect by summing the direct effect (*c*') and the indirect effect (*a* * *b*).

Mediation and bootstrapping

Although the parameter estimates for direct, indirect and total effects are easy to estimate for mediational models, correctly determining whether the *indirect effect* is statistically significantly different from 0 requires additional analytic attention. The indirect effect (*a* * *b*) is multiplicative. Even if the sampling distribution of both *a* and *b* are normally distributed, the sampling distribution of the *a* * *b* product is not necessarily normally distributed. Therefore, using the analytic standard error to determine the statistical significance of the *a* * *b* path may result in incorrect statistical inferences. Instead of trying to derive the standard error analytically, it is easy to *bootstrap* the sampling distribution around the *a* * *b* path. Bootstrapping is a resampling technique used to empirically derive the sampling distribution when an analytic solution is not feasible. Treating the sample (of size *n*) as the population, bootstrapping involves drawing repeated samples with replacement. The parameter estimates vary across samples. The variance of the parameter estimates provides an empirical estimate of the sampling variance; the standard deviation of the parameter estimates is an empirically derived standard error. However, because we believe that the sampling distribution of the indirect effect is not likely to be normal, we eschew standard errors and *p*-values (which assume that the distribution is normally distributed) in favour of empirically derived confidence intervals (CIs). To determine the 90% CI, we locate the 5th and 95th percentiles of the sampling distribution: those values become the upper and lower limits of our CI. For additional information about bootstrapping in SEM, we recommend Shrout and Bolger (2002), Preacher and Hays (2008) and MacKinnon and Fairchild (2009). Interestingly, recent research suggests that many of the bootstrap approaches have inflated Type I error rates (Yzerbyt et al., 2018). Therefore, Yzerbyt et al. (2018) suggest first examining the statistical significance of each of the paths separately. If both paths are statistically significant, then examine the magnitude and CI of the indirect effect using bootstrapping (Yzerbyt et al., 2018). Thus, in their approach, the tests of the individual components evaluate the statistical significance of the indirect effect whereas 'the confidence interval reveals its magnitude' (Yzerbyt et al., 2018, p. 942).

Mediation and causality

Mediation implies the existence of an underlying causal mechanism: the effect of a putative cause is transmitted through the mediator to the outcome variable (Mayer et al., 2014). However, since the advent of path analysis, controversy has surrounded the technique's causal aspirations (Wright, 1923). Recently, a great deal of methodological work has focused on whether and how researchers can make strong causal inferences from mediational models (Preacher, 2015; VanderWeele, 2015). Mediation analysis requires several fairly strong assumptions to attribute a causal interpretation to the indirect effect: (a) there are no omitted variables (confounders), (b) there is no measurement error in the predictor variable or mediators, (c) the functional form of the model is correct and (4) we have correctly modelled temporal precedence and the timing of measurement allows us to capture the mediation process (MacKinnon, 2008; MacKinnon et al., 2020). Because the term *mediation* implies an underlying causal mechanism, some researchers avoid using the term entirely, and instead refer only to the direct, indirect and total effects within path analytic (or structural equation) models with intermediate variables. To interpret estimates of direct and indirect effects causally does require fairly strong assumptions (VanderWeele, 2015). However, we choose to use the term *mediation* to describe models in which the effect of one variable is presumed to be transmitted through an intermediate variable to an outcome variable of interest, even when we fail to meet the strict assumptions of causal inference. We encourage our readers to read Volume 10 of this series, which is devoted to the topic of causal inference.

What are latent variables?

SEM is often referred to as a latent variable modelling technique (Hoyle, 2012). What are *latent variables*? The term *latent* means 'not directly observable'. Latent variables appear in a model but are not directly measured. We often teach our students a crude (but effective) rule of thumb for identifying whether a variable is latent or not. If the variable appears in the data file, it is observed. If the variable does not, it is latent. Bollen (2002) defines latent variables as variables for which there are no values (for at least some observations) in a given sample. Often, we use latent variables to model the hypothetically existing *constructs* of interest in a study that we cannot directly measure, such as peace, intelligence and apathy. However, not all latent variables are latent constructs. Bollen's definition of latent variables is broader and more inclusive than the definition of a latent construct. According to Bollen (2002), residuals (e.g. errors or measurement and disturbances) are also technically latent variables: they are not directly observed in a given sample. However, they are generally not latent constructs of substantive interest. We use the term *latent construct* to indicate a latent variable of substantive interest that assumes theoretical importance in a latent variable model. We use the term *latent variable* more broadly: a latent variable may be a latent construct of substantive interest, but it need not be. See Bollen (2002) for a far more nuanced discussion of the ways to define latent variables.

Measuring/modelling latent constructs

So, how can we model constructs that we cannot directly measure? In SEM, the existence of latent constructs is inferred or derived from the relationships among observed variables that measure the latent construct. To model latent constructs using reflective[ii] latent variables, we make two philosophical assumptions. First, we assume that the constructs are real, even if they cannot be directly measured (Borsboom, 2005; Borsboom et al., 2003; Cook & Campbell, 1979; Edwards & Bagozzi, 2000; Nunnally & Bernstein, 1994). Second, we assume that a latent construct has a causal relationship with its indicators (the observed variables that measure the construct of interest; Borsboom et al., 2003). In other words, the latent construct influences people's responses to the observed variables (or indicators; McCoach, Gable et al., 2013). Figure 4.3 illustrates this assumption. The circle represents the latent construct, the squares represent the observed variables that serve as indicators of the latent construct and the single-headed arrows represent the directional paths from the latent variable to the indicators. We can decompose the variance of each indicator into two parts: the variance that is explained by the latent construct and measurement error variance (Rhemtulla et al., 2020). This figure also illustrates one other implicit assumption of a standard, unidimensional factor model: we assume the correlations among the indicators are completely explained by the variance they share with the

[ii] You can also measure latent variables with formative or causal indicators. For a discussion of this approach, see Bollen and Diamantopoulos (2017) or Bollen and Bauldry (2011). However, in this book, we consider only latent variables with reflective indicators, which is by far the more common type of latent variable in the literature.

latent construct. In other words, we assume any covariances among the set of items are due to the latent construct.

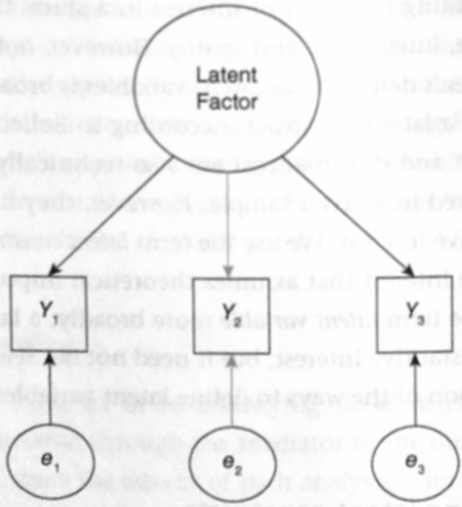

Figure 4.3 A simple unidimensional factor model

Factor analysis

To what extent do certain latent constructs explain the pattern of correlations/covariances in the observed variables? The goal of factor analysis is to determine the number of distinct constructs needed to account for the pattern of correlations among a set of measures (Fabrigar & Wegener, 2012). Factor analysis exploits the patterns of correlations among observed variables to make inferences about the existence and structure of latent constructs. Factor analysis provides information about which observed variables are most related to a given factor, as well as how these items relate to the other factors in the solution (Gorsuch, 1997).

Conceptually, standard factor analytic techniques assume that the correlations (covariances) among the observed variables can be explained by the factor structure. In other words, variance in the observed scores can be broken into two pieces: (1) variance that can be explained by the factor and (2) error variance (or uniqueness), which is the variance that is unique to the observed score and is not explained by the latent factor. Factors are the latent constructs of substantive interest that predict shared variance in the observed variables. Factor analysis yields estimates of the strength of the paths (measurement weights) from the latent factors to the indicators, the unique variance in each observed variable (the variance not explained by the factor) and the correlations among the latent variables of interest.

Types of factor analysis: exploratory and confirmatory factor analyses

The two most common factor analytic techniques are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Researchers commonly use EFA to reduce the number of elements from a larger number of observed variables to a smaller number of broader, more generalisable latent constructs (McCoach, Gable et al., 2013). Mathematically, EFA 'seeks the set of equations that maximise the multiple correlations of the factors to the items' (Gorsuch, 1997, p. 533).

One of the major methodological differences between EFA and CFA is the amount of information about the factor structure that is specified a priori. The factor structure represents the linkage between factors and indicators (i.e. which observed variables indicate which factor(s), how many factors are present in the data, etc.). EFA does not require a priori knowledge or specification of the factor structure. In contrast, in standard CFA, the researcher completely specifies the factor structure before undertaking the analysis. Based upon previous literature and experience, researchers clearly articulate the patterns of results they expect to find and then investigate whether and how well the data conform to the hypothesised structure.

CFA permits comparison of several rival models, allows researchers to reject specified models and provides a method to compare several competing models empirically. CFA has many advantages over EFA. These include (a) the ability to yield unique factorial solutions, (b) clearly defining a testable model, (c) assessments of the extent to which a hypothesised model fits the data, (d) information about how individual model parameters affect model fit, (e) the ability to test factorial invariance across groups (Marsh, 1987) and (f) the ability to compare and evaluate competing theoretical models empirically. Standard SEM techniques make extensive use of CFA in the development of the latent variables (i.e. measurement models). For the remainder of this book, we focus exclusively on CFA.

Measurement models versus structural models

In SEM, *measurement models* and *structural models* are conceptually distinct (Anderson & Gerbing, 1982, 1988). As previously mentioned, latent constructs represent theoretical constructs of interest that cannot be directly measured but that influence scores on the observed variables. To measure such latent constructs, we use multiple observed variables called *indicators*. The *measurement model* specifies the causal relations between the observed variables and the underlying latent variables (Anderson & Gerbing, 1982, 1988). The most common measurement model in SEM is a CFA model. For example, the unidimensional factor model in Figure 4.3 is also an example of a measurement model for a single latent construct.

In the standard conceptualisation of a measurement model, the only directional pathways are from latent variables to observed variables. Therefore, the latent constructs are *exogenous* variables. Generally, in SEM, all exogenous variables correlate with each other. In other words, the latent variables correlate with one another, but they do not predict one another.

The ***structural model*** is often the model of greatest interest: it specifies the causal or predictive pathways among the conceptual variables of interest. In such cases, the main purpose of the measurement model is to measure the theoretical constructs of interest both more completely and more accurately, using multiple indicators. Multiple indicators enable us to separate the 'true' variance of the latent variable from the measurement error that is inherent in each observed variable. The latent variables are measured without error, so the latent variable model generates unbiased estimates of the structural paths among the conceptual variables of interest. In Figure 4.4, we have reformulated our mediation model (from Figure 4.1) so growth mindset is no longer an observed variable; it is now a latent variable. The model now contains a latent variable for growth mindset as well as the overall structural model in which growth mindset predicts academic persistence, which in turn predicts academic achievement.

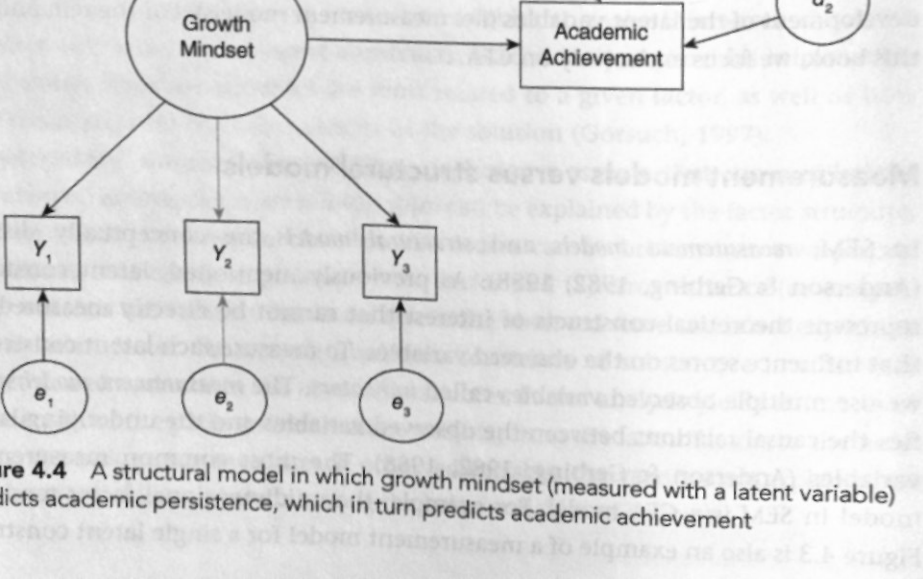

Figure 4.4 A structural model in which growth mindset (measured with a latent variable) predicts academic persistence, which in turn predicts academic achievement

Misspecified measurement models can lead to errors of inference in the structural part of the model. Therefore, the measurement model must be correctly specified and exhibit adequate fit prior to estimating the structural parameters (Anderson & Gerbing, 1982). We return to this point in Chapter 6, when we describe the model building process in SEM.

Disturbances and measurement errors

Generally, we refer to residuals for structural endogenous variables as *disturbances* and residuals for endogenous measurement variables as *measurement errors*. The disturbance variance represents the sum of all other causes of the endogenous *structural* variable that are *not* explicitly specified in the structural model. Similarly, error variance in the measurement model represents the sum of all other causes of the indicator variable that are *not* explained by the latent construct (factor). Note the difference between the use of *d*'s in Figure 4.1 and *e*'s in Figure 4.3. In either case, the total variance of any endogenous variable can be partitioned into two pieces: (1) the variance that is explained by its predictor variables and (2) the variance that is unexplained by the predictor variables. As in multiple regression, the proportion of explained variance in an endogenous variable is R^2. Therefore, the proportion of unexplained variance in an endogenous variable is $1 - R^2$.

Given that the latent variable has no inherent scale of its own, factor analytic results most commonly report the *standardised* path coefficients for a path from the latent variable to the indicator. These standardised path coefficients are also referred to as **measurement weights/pattern coefficients**, or *factor loadings* in CFA. In Figure 4.3, each of the pattern coefficients estimates the direct effect of the factor on the indicator variable.

The R^2 for a unidimensional indicator is simply the square of the standardised factor loading, and R^2 represents the proportion of variance in the indicator that is explained by the factor. For multidimensional indicators (where two or more factors predict a given indicator), we still can partition the variance in the indicator into the portion that is explained by the latent constructs and the portion that is unexplained by the latent constructs. The **proportion of variance unexplained by a factor** (or factors) is the measurement error variance for the indicator divided by the total variance of the indicator:

$$1 - R^2 = \frac{\textit{Measurement Error Variance}}{\textit{Total Variance}} \tag{4.1}$$

For example, suppose the total variance for an indicator is 100, and in a CFA measurement model, the error variance of that indicator is 20. The proportion of unexplained variance is 20/100 or .20. The R^2 for that indicator is 1 – proportion of unexplained variance, so $R^2 = 1 - .20$ or .80.

Estimation

The goal of estimation in SEM is to find the model parameter estimates that maximise the probability of observing the data. Finding parameter values for an overidentified model is iterative. The computer program repeatedly refines the parameter estimates to minimise the discrepancy between the model-implied variance–covariance matrix and the variance–covariance matrix (Brown, 2015). The ML estimates of the parameters minimise the discrepancy between the variance–covariance matrix and the model-implied variance–covariance matrix. 'ML aims to find the parameter values that make the data most likely' (Brown, 2015, p. 63). The structural equation model has *converged* when a unique set of parameter estimates minimise the difference between the model-implied and sample variance–covariance matrices. (*Note:* We find a unique set of parameter estimates for our specified model. However, that does not mean that our specified model is the only model to fit the data equally well!)

Model fit and hypothesis testing in SEM

In SEM, we fit our theoretical model to a variance–covariance matrix. The population covariance matrix represents bivariate relationships between the observed variables. The model parameters maximise the likelihood of obtaining the data, given the specified model. The estimated model parameters can then be used to generate the covariance matrix that is implied by the model. Ideally, the parameters in our model should be able to generate a *model-implied covariance matrix* that reproduces the population covariance matrix. The model-implied variance–covariance matrix provides important information about model–data fit. The more closely the parameters reproduce the covariance matrix, the better the 'fit' of the model. The fundamental statistical hypothesis undergirding SEM is H_0: $\Sigma = \Sigma(\theta)$. This null hypothesis states that the model-implied variance–covariance matrix is exactly equal to the population variance–covariance matrix. Here, Σ is the population covariance matrix, θ contains the set of parameters or system of equations and $\Sigma(\theta)$ is the model-implied covariance matrix (Paxton et al., 2011). The global fit function (F) measures the degree of discrepancy between the model-implied variance–covariance matrix and the actual variance–covariance matrix. The global fit function (F) tells us nothing about the predictive utility of the model. Instead, it is a function of the degree to which the model parameters are able to reproduce the covariance matrix.

How do we know if our model fits the data? Hypothesis testing in SEM departs from traditional tests of significance. In most statistical analyses, researchers test the null hypothesis that there is no relationship among a set of variables or that there are no statistically significant differences among a set of variables. Generally speaking, we want to reject the null hypothesis and conclude that there are statistically significant relationships or differences. In SEM, for the global test of model fit, the logic is reversed. We test the null hypothesis that the specified model exactly reproduces the population covariance matrix of observed variables (Bollen & Long, 1993). Assuming the data satisfy distributional assumptions (normality, etc.), the product of the asymptotic distribution of the fit function and the sample size minus 1 ($F * (N-1)$) is asymptotically distributed as chi-square (χ^2), with degrees of freedom equal to the degrees of freedom of the model. To evaluate exact global model fit, we compare the χ^2 of the specified model to the critical value for χ^2. Under the null hypothesis, the population covariance of observed variables equals the model-implied covariance matrix. Therefore, when the model χ^2 exceeds the critical value of χ^2, we reject the null hypothesis. Rejecting H_0 means the specified model does not adequately reproduce the covariance matrix, indicating less than perfect model fit.

However, this approach poses several problems. First, χ^2 is very sensitive to sample size. The larger the sample size, the more likely we are to reject the null hypothesis that the model fits the data. The χ^2 test rejects almost any model with a very large sample size if there is even a miniscule amount of model-data misfit. To correct for this problem, some researchers divide χ^2 by the model degrees of freedom. Ideally, the χ^2/df ratio should be less than 2. This does not really solve the problem though, as the degrees of freedom are related to model complexity and size, rather than sample size.

As mentioned earlier, ML estimation requires large sample sizes. This creates an obvious tension: we need large sample sizes in SEM, but large sample sizes provide high power to reject the null hypothesis that the model fits the data exactly. Because we hope to fail to reject the null hypothesis, having large sample sizes actually works against us.

Second, the χ^2 test is a test of exact (perfect fit). 'A perfect fit may be an inappropriate standard, and a high χ^2 may indicate what we already know – that H_0 holds approximately, not perfectly' (Bollen, 1989, p. 268). Knowing that the model-implied covariance matrix does not exactly fit the population covariance matrix provides no information about the degree to which the model does or does not fit the data. Scientific inquiry generally rewards parsimony and simplicity. Generally, models are simplifications of reality. In some sense, the goal of a model is to capture the essence of a system without completely recreating it. Therefore, it should come as no surprise that model-implied covariance matrices generally fail to exactly reproduce population covariance matrices.

Alternative measures of model fit

Because χ^2 is notoriously sensitive to sample size and the χ^2 test of model fit tends to be rejected with large sample sizes in SEM, SEM researchers have developed a variety of alternative global model fit measures to evaluate model–data fit (or misfit). The various alternative fit indices attempt to correct the problems that result from judging the fit of a model solely by examining χ^2.

There are two basic types of fit indices: (1) **absolute fit indices** and (2) **incremental fit indices**. Absolute fit indices evaluate the degree to which the specified model reproduces the sample data. Some of the more commonly used absolute fit indices include the root mean square error of approximation (RMSEA) and the standardised root mean square residual (SRMR). One of the most popular fit indices, the RMSEA, is a function of the degrees of freedom in the model, the χ^2 of the model and the sample size. Unlike the χ^2, the value of the RMSEA should not be influenced by the sample size (Raykov & Marcoulides, 2000). In addition, it is possible to compute a CI for the RMSEA. The width of the CI indicates the degree of uncertainty in the estimate in the RMSEA (Kenny et al., 2014). The SRMR represents a standardised summary measure of the model-implied covariance residuals. Covariance residuals are the differences between the observed covariances and the model-implied covariances (Kline, 1998). 'As the average discrepancy between the observed and the predicted covariances increases, so does the value of the SRMR' (Kline, 1998, p. 129). The RMSEA and the SRMR approach 0 as the fit of the model nears perfection. Hu and Bentler (1999) suggest that SRMR values of approximately .08 or below, and values of approximately .06 or below for the RMSEA indicate relatively good model fit.

Incremental fit indices measure the proportionate amount of improvement in fit when the specified model is compared with a nested baseline model (Hu & Bentler, 1999). Some of the most commonly used incremental fit indices include the non-normed fit index (NNFI), also known as the Tucker–Lewis Index (TLI) and the comparative fit index (CFI). Both indices approach 1.00 as the model–data fit improves, and the TLI can actually be greater than 1.00. Generally speaking, TLI and CFI values at or above .95 indicate relatively good fit between the hypothesised model and the data (Hu & Bentler, 1995, 1999) whereas values below .90 generally indicate less than satisfactory model fit. Many factors such as sample size, model complexity and the number of indicators can affect fit indices differentially (Gribbons & Hocevar, 1998; Kenny & McCoach, 2003; Kenny et al., 2014); therefore, it is best to examine more than one measure of fit when evaluating structural equation model. However, the vast array of fit indices can be overwhelming, so most researchers focus on and report only a few. We generally report χ^2, RMSEA, the SRMR, the CFI and the TLI.

Model comparison

Chi-Square difference test

The χ^2 difference test compares the model fit of two hierarchically nested models. Two models are hierarchical (or nested) models if one model is a subset of the other. For example, if a path is removed or added between two variables, the two models are hierarchical (or nested) models (Kline, 2015). However, models that simultaneously free one or more parameters while constraining one or more previously freed parameters are not nested. For the χ^2 difference test, we subtract the χ^2 of the more complex model (χ_2^2) from the χ^2 of the simpler model (χ_1^2). We then subtract the degrees of freedom of the more complex model (df_2) from the degrees of freedom for the more parsimonious model (df_1). We compare this χ^2 difference ($\chi_1^2 - \chi_2^2$) to the critical value of χ^2 with $df_1 - df_2$ degrees of freedom. If this value is greater than the critical value of χ^2 with $df_1 - df_2$ degrees of freedom, we conclude that deleting the paths in question has statistically significantly worsened the fit of the model. If the value of $\chi_1^2 - \chi_2^2$ is less than the critical value of χ^2 with $df_1 - df_2$ degrees of freedom, then we conclude that deleting the paths has not statistically significantly worsened the fit of the model. When deleting paths does not worsen the fit of the model, we choose the more parsimonious model (the one that has fewer paths and more degrees of freedom) as the better model. The χ^2 difference test can only be used to compare hierarchically related models. If observed variables are added or removed from the model (i.e. if the observed variance–covariance matrix changes), the models are not hierarchical models. It is inappropriate to use the χ^2 difference test to compare models that have different numbers of variables or different sample sizes.

Because sample size affects χ^2, sample size also affects the χ^2 difference test. Small differences between the observed and model-implied variance–covariance matrices can produce a large χ^2 when the sample size is very large. Likewise, all else being equal, we are more likely to observe statistically significant χ^2 differences between two hierarchically nested models in a large sample than in a small sample. Therefore, any results should be viewed as a function of the power of the test as well as a test of the competing models.

Fitting multiple models

Generally, SEM specifies an a priori model, based on previous literature and substantive hypotheses. Unlike traditional statistical techniques, in SEM, it is common to evaluate several models before adopting a final model. Sometimes, after fitting the initial model, a researcher might wish to change certain aspects of the model, a process called **respecification**. There are at least three distinct reasons for estimating multiple structural equation models.

1 Theorists seek the most parsimonious explanation for a given phenomenon. The initial model includes all possible parameters. Subsequent models eliminate unnecessary (non-statistically significant) parameters, a process that is sometimes called *trimming* the model. They test the fit of the new more parsimonious model (with greater degrees of freedom) against the original model using the χ^2 difference test. This practice is far more defensible when eliminating paths that are conceptually expected to be 0 than when model trimming is conducted for purely empirical reasons (i.e. all paths that are non-statistically significant are omitted for purely empirical reasons; Kline, 2015).

2 The researcher compares two or more competing theoretical models. Using SEM, the researcher(s) specify the competing models a priori and then compare the models to determine which model appears to better fit the data.

3 The initial model exhibits poor fit. Subsequent models seek to find a model that provides better fit to the data. Purely empirically motivated model modifications lead down a treacherous path, as we discuss next.

Model modification and exploratory analyses

Modification indices

If the model does not exhibit adequate fit, how should the researcher proceed? SEM output may include *modification indices* (sometimes called Lagrange multiplier tests). The modification index for a parameter is the expected drop in χ^2 that would result from freely estimating a particular parameter. Remember, χ^2 drops as we add parameters to our model and lower χ^2 values indicate better fit. If we add a parameter to our model, the χ^2 needs to decrease by at least 3.84 points to be statistically significant at the .05 level. Therefore, some researchers request all modification indices above 4. This provides a list of parameters that could be added to the model that would result in a statistically significant decrease in χ^2. Modification indices are univariate. Therefore, adding two parameters simultaneously would not necessarily result in a change in χ^2 equal to the sum of the two modification indices.

The modification indices suggest which parameters might be added to the model to improve model fit. Parameters with larger modification index values result in larger decreases in χ^2, resulting in greater improvements in model fit. Thus, it can be tempting to use these modification indices to make changes to improve the fit of the model. Proceed very cautiously! Although some suggested model modifications may be conceptually consistent with the research hypotheses, other model modifications may make no conceptual sense. Sometimes the modifications suggested by the SEM program are downright illogical and indefensible. Second, making changes based on modification indices (or model fit more generally) capitalises on chance idiosyncrasies of the sample data and may not be replicable in a new sample. Respecification of structural equation models should be guided by theory, not simply by a desire to improve measures of model fit. A good analyst uses modification indices very cautiously (if at all) and reports and substantively defends each modification.

One of the most common and controversial practices in SEM is model modification. When the model which was specified a priori does not exhibit good fit to the data, the temptation to modify the model to achieve better fit can be irresistible. SEM software programs provide suggested model modifications, based solely on statistical criteria. The researcher is then left to determine what, if any, model modifications are warranted. Although using empirical data to guide decision-making may be helpful for 'simple' modifications, it does not tend to inform 'major changes in structure', and some indications for change may be 'nonsensical' (Bollen, 1989, p. 296). Moreover, when we use the same set of data to both develop a model and evaluate its fit, we undermine the confirmatory nature of our analyses (Breckler, 1990). Further, making modifications based on the desire to improve model capitalises on sampling error; such modifications are unlikely to lead to the true model (Kline, 2015; MacCullum, 1986). If the initial model is incorrect, it is unlikely that specification searches will result in the correct model (Kelloway, 1995). Therefore, we advise against blindly following the brutally empirical suggestions of model modification indices. Models with more parameters may fit the data better simply because of chance fluctuations in the sample data. In essence, we can overfit a model to a set of data, and such models will not replicate well with a new sample. 'A model cannot be supported by a finding of good fit to the data when that model has been modified so as to improve its fit to that same data' (MacCallum, 2001, p. 129). Therefore, replication, not modification, provides the best path to enlightenment in SEM.

Model fit versus model prediction

A model may exhibit adequate fit, and yet do a poor job of predicting the criterion variable of interest. In fact, a model with no statistically significant parameters can fit well (Kelloway, 1995), whereas a 'poor' fitting model may explain a large amount of the variance in the outcome of interest. In fact, models with highly reliable manifest indicators tend to exhibit worse fit than models with less reliable indicators (Browne et al., 2002). Many researchers who would never neglect to report the R^2 for a multiple regression analysis seem to overlook the importance of reporting similar measures of variance explained within SEM. Because SEM places a great deal of emphasis on model fit, some researchers lose sight of the fact that a good fitting model may explain very little variability in the variable(s) of interest. To assess **model prediction** for a given endogenous (dependent) variable, we compute the proportion of variance in the variable that is explained by the model. The ratio of the variance of the disturbance (or error) to the total observed variance represents the proportion of unexplained variance in the endogenous variable. Therefore, R^2 is simply 1 minus that ratio (Kline, 2015). Determining the variance explained in non-recursive models is more complex. See Bentler and Raykov (2000) for details on calculating of R^2 for non-recursive models.

When SEM gives you inadmissible results

In addition to examining the parameter estimates, the tests of significance and the fit indices, it is very important to examine several other areas of the output to ensure that the program ran correctly. The variances of the error terms and the disturbances should be positive and statistically significant. As in multiple regression, the standardised path coefficients generally fall between –1.00 and +1.00. Further, the standardised error terms and disturbances should fall in the range of 0.00 to 1.00. Negative error variances and correlations above 1 are called **Heywood cases**, and they indicate the presence of an **inadmissible solution**. Heywood cases can be caused by specification errors, outliers that distort the solution, a combination of small sample sizes and only one or two indicators per factor, or extremely high or low population correlations that result in empirical underidentification (Kline, 2015).

Additionally, the SEM program may fail to converge in the allotted number of iterations. Lack of convergence indicates that the algorithm failed to produce an ML solution that minimises the distance between the observed and model-implied covariance matrices. Again, when this happens, the output should not be trusted. Requiring very large or infinite numbers of iterations can be signs of a problem such as an underidentified model, an empirically underidentified model, bad start values, extreme multicollinearity, a tolerance value approaching 0 or other specification error (Kline, 2015). If the program fails to converge, inspect the output for possible errors or clues to the reason for the non-convergence, respecify the model to address the problem, and run the SEM again. It is never advisable to interpret output that contains any Heywood cases, non-convergent or inadmissible solutions.

Alternative models and equivalent models

In traditional SEM, we specify a particular model a priori, but our hypothesised model is statistically equivalent to a myriad of models. Two models are equivalent if they reproduce the same set of model-implied covariance (and other moment) matrices (Hershberger, 2006; Raykov & Penov, 1999; Tomarken & Waller, 2003). **Equivalent models** have different causal structures but produce identical fit to the data (Hershberger, 2006). Equivalent models produce identical values for the discrepancy between the model-implied matrix and the observed matrix; therefore, they result in identical values for model χ^2 and model fit indices. Perhaps the simplest example of model equivalence is to reverse the causal paths in a path analytic diagram. For example, specifying that $X \rightarrow Y \rightarrow Z$ is equivalent to specifying that $Z \rightarrow Y \rightarrow X$. For complex models, there are often dozens (or even hundreds!) of functionally equivalent models that the researcher has not tested. Hershberger (2006), Lee and Hershberger (1990) and Stelzl (1986) demonstrate rules for generating multiple equivalent models. Even when a model fits the data well, any statistically equivalent models would fit the data equally well (Tomarken & Waller, 2003, 2005). Equivalent models can lead to substantially different theoretical or substantive conclusions (Hershberger, 2006; MacCallum et al., 1993; Tomarken & Waller, 2005). Unfortunately, researchers often fail to recognise the existence of equivalent models or consider equivalent models when interpreting the results of their research (MacCallum et al., 1993).

In addition, an untested model may provide even better fit to the data than the researcher's hypothesised model, and there is no fail-safe method to protect against this possibility. Researchers who test an assortment of plausible competing models can bolster their argument for a particular model. However, because the number of rival alternative models may be virtually limitless, testing multiple competing models does not eliminate the possibility that an untested model may provide better fit to the data than does the researcher's model. Therefore, any specified model is a tentative explanation and is subject to future disconfirmation (McCoach et al., 2007).

A word of caution

SEM is a powerful data analytic technique, but it is not magic. No matter how appealing and elegant SEM may be, it is a data analytic technique, and as such, it is incapable of resolving problems in theory or design (McCoach et al., 2007). The adage 'correlation does not imply causation' applies to SEM as well. Although SEM may appear to imply or specify causal relations among variables, causality is an assumption rather than a consequence of SEM (Brannick, 1995). Wright's original description of the technique still holds true today:

> The method of path coefficients does not furnish general formulae for deducing causal relations from knowledge of correlations and has never been claimed to do so. It does, however, within certain limitations, give a method of working out the logical consequences of a hypothesis as to the causal relations in a system of correlated variables. The results are obtained by a combination of the knowledge of the correlations with whatever knowledge may be possessed, or whatever hypothesis it is desired to test, as to causal relations. (Wright, 1923, p. 254)

Using SEM allows us to ascertain whether a hypothesised causal structure is consistent or inconsistent with the data; however, causal inferences ultimately depend 'on criteria that are separate from that analytic system' (Kazantzis et al., 2001, p. 1080).

SEM can polarise researchers: the technique has been both demonised and canonised (Meehl & Waller, 2002). When applied and interpreted correctly, SEM is an invaluable tool that helps us to make sense of the complexities of our world, and SEM offers several advantages over traditional statistical techniques. However, SEM does not replace the need for good design and sound judgement. Causal inference in SEM requires the same assumptions as any other methods, and the ability to make causal inferences rests firmly on the design of the study. Strong designs allow for stronger causal inferences; weak designs lead to weak causal inferences. Just as 'no amount of sophisticated analyses can strengthen the inference obtainable from a weak design' (Kelloway, 1995, p. 216), no analytic method can replace the need for critical appraisal and common sense.

SEM is both an art and a science. Because structural equation models are so open to modification, SEM allows for a great deal of artistic license on the part of the analyst. SEM allows researchers a great deal of flexibility and control over their analyses, which provides opportunities for both innovation and manipulation. It is this freedom that makes SEM so powerful and so appealing, but also so prone to misuse. With this flexibility comes great responsibility. We must build our models thoughtfully, describe our model building process fastidiously and interpret the results of our SEM analyses cautiously. Producers and consumers of structural equation model should realistically evaluate the strengths and limitations of this technique and should interpret the results of SEM analyses accordingly. In Chapter 6, we provide concrete recommendations for defensible model building processes. However, before doing so, Chapter 5 delves into the important foundational topics of SEM specification and identification.

Chapter Summary

- Structural equation modelling (SEM) refers to a family of techniques, including (but not limited to) path analysis, confirmatory factor analysis, structural regression models, autoregressive models and latent change models.
- SEM allows researchers to distinguish between observed and latent variables and to explicitly model both types of variables.
- Using latent variables in SEM accounts for potential errors of measurement, allowing researchers to explicitly account for (and model) measurement error (Raykov & Marcoulides, 2000). The ability to separate measurement error or 'error variance' from 'true variance' is one of the reasons that SEM provides such powerful analyses.

- SEM is a regression-based technique. As such, it rests upon four key assumptions and requirements necessary for any regression-based analyses: (1) normality, (2) linearity, (3) independence and (4) adequate variability.
- The basic building block of any structural equation model is the variance–covariance matrix.
- Models with more parameters may fit the data better simply because of chance fluctuations in the sample data. It is possible to overfit a model to a set of data, and such models will not replicate well with a new sample.
- SEM allows researchers a great deal of flexibility and control over their analyses, which provides opportunities for both innovation and manipulation. In SEM, we must build models thoughtfully, describe the model building process fastidiously and interpret the results of analyses cautiously.

Further Reading

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

This article discusses the definition and use of latent variables in psychology and social science research.

Kline, R. B. (2015). *Principles and practices of structural equation modeling* (4th ed.). Guilford Press.

This book is an oft-cited introductory guide to structural equation modelling. It provides a non-technical introduction to structural equation modelling as well as several other topics including multilevel structural equation modelling, growth curve analysis and mean and covariance structure analysis.

5

SPECIFICATION AND IDENTIFICATION OF STRUCTURAL EQUATION MODELS

Chapter Overview

In Chapter 4, we introduced SEM as a technique for analysing **systems of equations** in which path diagrams pictorially represent systems of equations. This chapter provides an overview of model identification criteria and explicitly links path diagrams to the structural equations that they specify. After explaining the link between path diagrams and structural equations, we demonstrate how to use Wright's rules to derive the **model-implied correlation/covariance** matrix for a given SEM. Appendix 2 provides a more technical introduction to the link between path diagrams and structural equations. Appendices 3 and 4 provide even greater detail on Wright's rules. Appendix 3 demonstrates Wright's standardised tracing rules. Appendix 4 discusses Wrights unstandardised rules and covariance algebra.

SEM involves solving of a set of simultaneous equations in which the known values are a function of the unknown parameters (Kenny & Milan, 2012). In Chapters 4 to 6, our known values consist of the observed variances and covariances because we limit our discussion to models that do not include means or mean structure. However, in Chapter 7, we introduce means and mean structure into our SEMs.

To generate unique estimates for all these parameters, the SEM model must be identified (Kline, 2015). Identification involves demonstrating 'that the unknown parameters are functions only of the identified parameters and that these functions lead to unique solutions' (Bollen, 1989, p. 88). If all parameters in the model are uniquely identified, then the model itself is identified. We provide a brief introduction to identification rules for **recursive structural equation models.** Recursive structural equation models have no feedback loops and no correlated disturbances. Therefore, any variable (*Y*) cannot both be a predictor of and predicted by another variable (*X*). The rules of identification for **non-recursive structural equation models** are far more complicated than the rules for recursive models. Given the introductory nature of this text, we present identification rules for recursive models only. However, readers who are interested in learning more about non-recursive models should read Paxton et al. (2011), which provides a very approachable introduction to non-recursive models. For a fuller discussion of identification issues, see Rigdon (1995), Kline, 2015, Kenny and Milan (2012) or Steiger (2002).

Computing degrees of freedom in SEM

As mentioned in Chapter 4, the covariance matrix serves as a sufficient statistic for standard SEM: raw data are not necessary. Instead, it is possible to estimate SEM parameters using the covariance matrix. In statistical analyses such as ANOVA and regression, the degrees of freedom are a function of the sample size. In contrast, in SEM, the number of parameters that we can freely estimate is limited by the number of unique elements in the variance–covariance matrix (or the variance–covariance matrix plus means for models that include means). We cannot estimate more parameters than there are unique elements in the variance–covariance matrix, no matter how large our sample size is!

Let's count the number of unique elements for the small variance–covariance matrix shown in Equation (5.1). There are six unique elements: three diagonal elements (the three variances) and three unique off-diagonal elements (the three unique covariances: cov_{12}, cov_{13} and cov_{23}). (The covariances below the diagonal are identical to the covariances above the diagonal; that is, the covariance between variables 1 and 3 is identical to the covariance between variables 3 and 1, so we count each of these covariances only once.)

$$\begin{bmatrix} \text{var}_1 & \text{cov}_{12} & \text{cov}_{13} \\ \text{cov}_{21} & \text{var}_2 & \text{cov}_{23} \\ \text{cov}_{31} & \text{cov}_{32} & \text{var}_3 \end{bmatrix} \tag{5.1}$$

The number of unique elements in the variance–covariance matrix is the number of *knowns* (the information that we *know* before we begin our analyses).

If there are many variables in the model, counting the number of unique elements in the variance–covariance matrix is tedious. Luckily, there is an easy formula to compute the number of unique elements in the variance–covariance matrix. The number of unique elements in the variance–covariance matrix (the knowns) equals

$$\text{Knowns} = \frac{v(v+1)}{2} \tag{5.2}$$

where v = the number of observed variables in the variance–covariance matrix. Therefore, if there are 20 observed variables in the variance–covariance matrix, then the number of knowns equals $\frac{20(20+1)}{2} = \frac{20 \cdot 21}{2} = 210$. A slight modification to the formula calculates the number of off-diagonal elements of the variance–covariance matrix (i.e. the number of unique **correlations** in a correlation matrix): $v(v-1)/2$, where v is (still) the number of observed variables.

The number of *knowns* places an upper limit on the number of possible *unknowns*, which are the freely estimated parameters in the model. In SEM, we estimate several different types of parameters: exogenous variances and endogenous variances (which can be either disturbances or measurement errors), paths and **covariances/correlations**. For the number of parameters in the model (unknowns), we count the number of freely estimated variances, paths and covariances/correlations. The degrees of freedom in SEM equal the number of knowns (unique elements of the variance–covariance matrix) minus the number of unknowns (freely estimated parameters). A model has positive degrees of freedom if the model contains fewer parameters than there are unique elements in the variance–covariance matrix. Measures of model fit (chi-square, RMSEA, CFI, etc. are only available for models with positive degrees of freedom).

Figure 5.1 depicts a simple path model with four observed variables: (1) parental expectations, (2) growth mindset, (3) academic persistence and (4) academic achievement. Of course, these constructs could be measured using latent variables, and that would be preferable. However, to start simply, we demonstrate tracing rules with a path model. The model estimates paths from parental expectations to growth mindset, academic persistence and academic achievement, from growth mindset to academic persistence and from academic persistence to academic achievement.

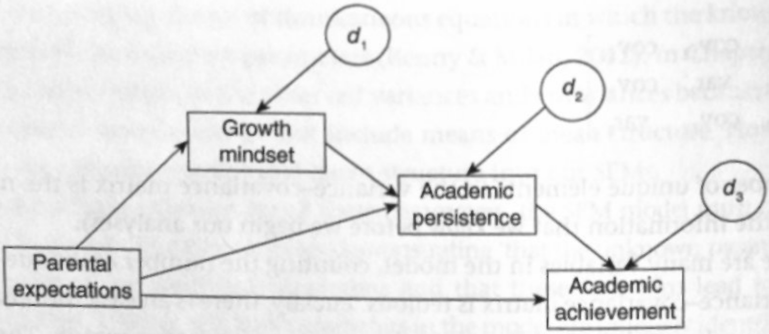

Figure 5.1 A simple path model with four observed variables

How many degrees of freedom does this model contain? First, we count the freely estimated parameters. We estimate variances for every exogenous variable and disturbance variances for every endogenous variable in the model. The model in Figure 5.1 estimates one exogenous variance (for parental expectations) and three disturbance variances (for growth mindset, academic persistence and academic achievement). The model also includes five freely estimated paths and zero covariances. Therefore, the number of freely estimated parameters (unknowns) equals 9 (5 paths + 1 exogenous variance + 3 disturbance variances). There are four observed variables, so the number of unique elements in the covariance matrix equals 4 * 5/2, or 10. The degrees of freedom for the model equals the number of knowns (10) minus the number of unknowns (9), or 1 *df*. Why is there 1 *df*? Our model contains no direct effect of growth mindset to academic achievement. There is no path from growth mindset to academic achievement, so that path is constrained to 0. By eliminating that path, in our hypothesised model, the effect of growth mindset on academic achievement is completely mediated by academic persistence. For our model to fit the data, the correlation between growth mindset and academic achievement must be completely explained by their mutual relationships with academic persistence. If this were not true, the hypothesised model (depicted above) would fit more poorly than the model that includes that path. Because this model has only 1 *df*, we know that the χ^2 of the model is completely attributable to the misfit due to the elimination of the path (or direct linkage) from growth mindset to academic persistence.

When the number of knowns equals the number of unknowns, the model is said to be just-identified. A just-identified model contains as many knowns as unknowns, so the parameter estimates can always perfectly reproduce the variance–covariance matrix. Thus, the just-identified model 'fits' the variance–covariance matrix perfectly. Just-identified models always have 0 *df*. Adding a path (direct effect) from growth mindset to academic achievement produces a just-identified (fully saturated) model. For all just-identified models, both the χ^2 and *df* are 0.

In fact, all multiple regression models are actually just-identified path models, so they have 0 *df*. The knowns are the number of observed variables, which is the sum of the predictors and the outcome variable. The unknowns are the regression coefficients, the exogenous variances for the predictors, the residual variance for the outcome variable and the covariances among all of the exogenous variables (the predictors). Because we allow all exogenous variables to correlate with each other, the number of freely estimated parameters is exactly equal to the number of unique elements of the variance–covariance matrix.

If the specified model requires estimating more parameters than there are unique pieces of information in the variance–covariance matrix, the model has negative degrees of freedom and is underidentified. It is not possible to solve the set of structural equations for underidentified models because there are more unknowns than knowns. Just as it is not possible to find a unique solution to the equation $x + y = 10$ because the number of unknowns is greater than the number of knowns, it is not possible to uniquely identify all of the parameters in a model with negative degrees of freedom. Having non-negative degrees of freedom is a necessary (but not sufficient) condition for model identification: models with positive degrees of freedom can still be underidentified. The problem of underidentification is theoretical rather than statistical (Heise, 1975); it is not data dependent. (*Empirical* underidentification, on the other hand, is data dependent. See Kenny & Milan, 2012, for more information about empirical underidentification.) Therefore, it is important to evaluate whether or not the structural equation models of interest are identified or identifiable during the design phase of the study, prior to data collection.

Luckily, although some of the identification rules for SEM are quite complex, the identification rules for recursive path models are actually quite simple. Recursive path/structural models with non-negative degrees of freedom are always identified. However, non-recursive structural equation models with positive degrees of freedom may not be identified. For more information on the identification of non-recursive models, see Berry (1984), Nagase and Kano (2017) and Paxton et al. (2011). The rules of identification also become more complex for measurement (factor) models, as we shall soon see.

Assuming that the model is not unidentified (inestimable) due to other problems in the specification, a recursive path model with positive degrees of freedom is *overidentified.* An overidentified model uses a smaller number of parameters to estimate all elements of the variance–covariance matrix, resulting in some discrepancy between the available variance–covariance matrix and the parameters to be estimated (Kenny & Milan, 2012).

In the case of overidentification, there is more than one way to estimate one or more of the parameters in the system of equations (Kenny, 2004). An overidentified model is more parsimonious than a just-identified model: it attempts to reproduce all the elements of the variance–covariance matrix with fewer parameters. As such, it is a simplification of reality. However, some level of detail or information is lost in that process. In such a scenario, we favour the solution that produces parameter estimates that maximise the likelihood of observing our data.

SEM model fit is an indication of the degree to which our simplified model reproduces (or fails to reproduce) the variance–covariance matrix (Kenny & Milan, 2012). Measures of model fit (e.g. the χ^2 test) are available for overidentified models; thus, it is possible to evaluate the fit of an overidentified model and test it against other competing models. In fact, it is only possible to examine model fit for models that are overidentified.

As we discussed in Chapter 4, one of the great advantages of using SEM is the ability to incorporate latent variables. The path models that we just described specified structural relationships among variables. However, thus far, our simple path/mediation models have contained observed variables (but not latent variables). Next, we introduce the specification and identification of measurement models, which specify the (causal) relationships among latent and observed variables. Afterwards, we demonstrate the integration of path models and measurement models to estimate hybrid structural equation models with latent variables.

Model specification of measurement (CFA) models

Path diagrams, visual displays of latent variable models, are the most intuitive way to conceptualise measurement models. Figure 5.2 depicts a measurement model for math and reading ability as a standard CFA model. Recall that in a path diagram, rectangles denote the indicators, or the observed variables, and circles represent latent variables. Each indicator is predicted by both a latent variable and an error. The small circles with δ's represent the measurement errors, or the residuals. The curved line between the two latent variables indicates a covariance between the two exogenous latent variables.

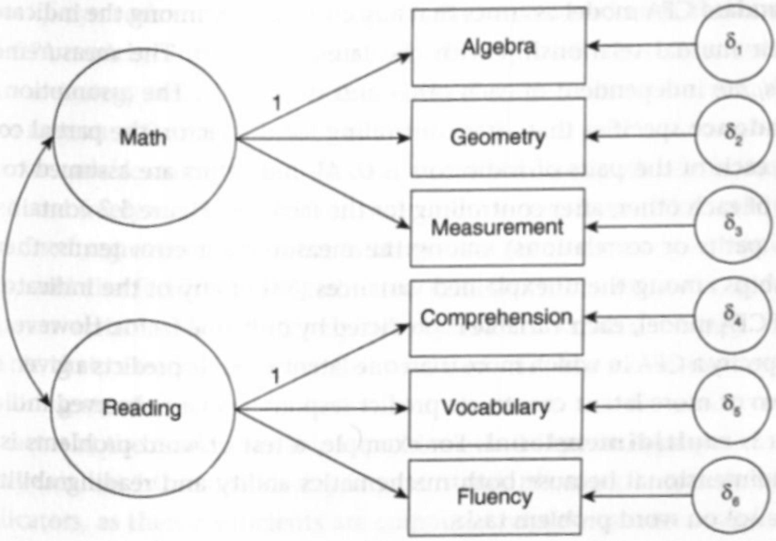

Figure 5.2 A standard confirmatory factor analysis (measurement) model for math and reading ability

In Figure 5.2, paths connect each of the factors to the three indicators of those factors. Just as in multiple regression and path models, these paths indicate a direct effect from the factor to the indicator (observed variable) *after controlling for* any other variables that also have direct effects on the indicator.

In a standard CFA model, each observed variable is an indicator of only one latent variable, and each observed variable is predicted by both the latent variable and an error. Two sources influence a given indicator – the factor (*F*) and the measurement error term (δ), which encompasses all other unique sources of variance. In other words, a person's response to an item is determined partially as a function of his or her standing on the factor (latent variable) and partially as a function of error, noise or other variables that are not part of the model. The variance in the indicator (the observed variable) consists of two pieces: (1) the variance that can be explained by the latent variable (the factor variance) and (2) the variance that is not explained by the latent variable (δ, the measurement error variance). We distinguish between **measurement error**, residual (error) variance in the measurement model, and disturbance variance, residual variance in the structural model.

A standard CFA model assumes that each factor is **unidimensional**. Conceptually, imagine the attribute being measured by a unidimensional factor falling on a straight line: people who are high on the factor possess more of the attribute and people who are low on the factor possess less of it. Unidimensionality indicates that the statistical dependence among a set of indicators is captured by a single latent construct (Crocker & Algina, 1986; McCoach, Gable et al., 2013).

The standard CFA model assumes that any correlations among the indicators result from their mutual relationship with the latent factor(s). The measurement error terms (δ's) are independent of each other and the factors. The assumption of **local independence** specifies that after controlling for the factor, the partial correlation between each of the pairs of indicators is 0. All indicators are assumed to be independent of each other, after controlling for the factor(s). Figure 5.2 contains no linkages (i.e. paths or correlations) among the measurement error terms: there are no relationships among the unexplained variances (δ's) of any of the indicators. In the standard CFA model, each variable is predicted by only one factor. However, it is possible to specify a CFA in which more than one latent variable predicts a given indicator. When two or more latent constructs predict responses on an observed indicator, the indicator is **multidimensional.** For example, a test of word problems is likely to be multidimensional because both mathematics ability and reading ability predict performance on word problem tasks.

Figure 5.2 contains a direct path from math to algebra but no direct path from reading to algebra, which means the direct path from reading to algebra is constrained to be (fixed at) 0. In other words, the model assumes that there is no direct effect of reading on algebra, after controlling for math. This does *not* mean that reading is unrelated to the indicators of mathematics achievement. Rather, the model specifies that the relationship between the reading factor and algebra is indirect: it is a function of the relationship between the reading and math factors and the path from the math factor to the algebra indicator. Because the model contains no direct effect of reading on algebra after controlling for math achievement, the standardised path coefficient from math to algebra is also the model-implied correlation between the math factor and algebra. Squaring the standardised path coefficient from math to algebra computes the proportion of variance in the algebra indicator that is explained by the math factor (R^2). The proportion of variance in the indicator that is not explained by the factor is $1 - R^2$, which is also the error variance of the indicator divided by the total variance of the indicator.

Model identification: measurement models/CFA

Standard CFA estimates parameters for the paths from the latent factors to the observed variables (the factor loadings), the variances of the latent variables, the variances of the measurement errors and the covariances (or correlations) among the latent variables. The factors have no inherent scales because they are latent variables; they are not actually measured or observed. Therefore, to identify the CFA model, we must *scale* the latent variable. Scaling the latent variable provides an anchor to and meaning for the metric of the latent variable. Two common options for scaling the latent variable are the **fixed factor variance strategy** or the **marker variable strategy**.

The *fixed factor variance strategy* constrains the variance of each factor to 1.0. In standard CFA models, the factor's mean is constrained to 0. Therefore, the fixed factor variance strategy results in a standardised solution: each latent variable in the model has a mean of 0 and a variance of 1.

The *marker variable strategy* constrains one unstandardised path coefficient for each factor to 1 and freely estimates the variance of the latent factor. The variable whose unstandardised regression coefficient is fixed to 1 becomes the *marker variable* for the factor, and the factor's freely estimated variance is scaled in the same metric as the marker variable. Figure 5.2 employs the marker variable strategy. Generally, any variable that is reasonably strongly correlated with the factor can be a marker variable. Using an observed variable that is uncorrelated or only weakly correlated with the factor as the marker variable is problematic. Why? Fixing a path coefficient that is very small at 1.0 results in very large unstandardised path coefficients for the other indicators, as their coefficients are computed relative to the coefficient of the poor indicator. Therefore, it is advisable to select one of the strongest indicators as the marker variable. If the variables are measured in different metrics, it is helpful to choose a variable with the most interpretable metric to be the marker variable, given that the latent factor is scaled in the metric of the marker variable.

In standard, single-group CFA, the two scaling methods (fixed factor variance and marker variable) result in statistically equivalent models. The marker variable strategy is the default approach in many statistical software packages (e.g. Mplus, AMOS, lavaan). Furthermore, it is common to use the marker variable approach when conducting multiple-groups CFA to assess the invariance of the model factor structure across different subsets of a sample. So, how does the latent variable scaling method impact the estimated parameters in a CFA model? The standardised parameter estimates are identical, regardless of the scaling technique employed. However, the unstandardised parameter estimates differ across scaling techniques. Using the fixed factor variance strategy, the unstandardised parameter estimates are identical to the standardised parameter estimates. The marker variable approach scales the unstandardised results in the metric of the marker variable for each factor.

Identification in CFA models

Freely estimated parameters in CFA

The marker variable strategy constrains one path coefficient (factor loading) per factor to be 1.00 and freely estimates the remaining unstandardised path coefficients. Using the marker variable strategy, we estimate a factor variance for each of our latent variables and an error variance for each of our observed variables. Generally, standard CFA models allow all factors (which are exogenous latent variables) to be intercorrelated;

estimating inter-factor covariances (correlations) for all factors. If we allow any measurement errors to be correlated (covary), then we must count those correlations (covariances) as estimated parameters as well. The number of unknowns (parameters to be estimated) equals the sum of the freely estimated paths, factor variances and covariances, and error variances and covariances. Recall, the number of knowns equals the number of unique elements in the variance–covariance matrix, which can be calculated using the formula: $v(v + 1)/2$, where v is the number of observed variables. The degrees of freedom for a given model equals the number of knowns minus the number of unknowns.

How many degrees of freedom does a single-factor model with three indicators have? A single-factor model with three indicators estimates six parameters (unknowns): one factor variance, two paths and three error variances. With three observed variables, the number of unique variances and covariances (the knowns) equals $3 * 4/2 = 6$. The number of knowns (6) equals the number of unknowns (6). Thus, a single-factor model with three indicators is just-identified: it has 0 *df*. A single-factor model with four or more indicators is overidentified. For example, a standard single-factor model with four indicators contains 2 *df*. Why? This model contains eight unknowns (parameters to be estimated): one factor variance, three path/pattern coefficients and four error variances. The number of knowns equals the number of unique elements in the variance–covariance matrix, which is $4 * 5/2$, or 10. There are $10 - 8 = 2$ *df*. See if you can explain why a single-factor model with five indicators has 5 *df*.[1]

How many degrees of freedom does the model in Figure 5.2 contain? Using the fixed factor variance strategy, we estimate six paths, one covariance and six measurement error variances, for a total of 13 parameters. Using the marker variable strategy, we estimate two factor variances, four paths, one covariance and six measurement error variances. Using either strategy, we estimate 13 parameters (unknowns). There are six observed variables in the model, so the variance–covariance matrix contains $6 * 7/2 = 21$ unique elements. Therefore, the model in Figure 5.2 contains 8 *df* ($21 - 13 = 8$). Where are these 8 *df*? The model above reproduces 21 variances and covariances (six variances for the six observed variables and all of their covariances) using only 13 freely estimated parameters. Although the source of the degrees of freedom may seem less obvious in the measurement model, the logic is the same: a variance–covariance matrix contains a linkage between every observed variable. Using a fixed factor variance strategy, we estimate all paths and constrain the latent variances to 1. When fixing the factor variances to 1, the number of freely estimated paths and correlations in the measurement model may not exceed the number of unique correlations in the **covariance/correlation matrix**.

[1] We are estimating one factor variance, four paths and five error variances, so the total number of unknowns = 10. The number of knowns = $5 * 6/2 = 15$. $15 - 10 = 5$. Therefore, a single-factor model with five indicators has 5 *df*.

How many observed variables are needed to adequately measure each latent variable? This is a very complex and nuanced issue. Using three or more observed variables is technically sufficient to estimate a single latent variable (factor). A standard one-factor model with three observed variables is just-identified. With four or more observed variables, the single-factor model is overidentified. For models with multiple factors, as few as two observed variables per factor might be technically adequate. However, from a theoretical standpoint, adequately measuring the latent variable of interest may require more indicators than are technically necessary. In general, the more abstract and loosely defined a construct is, the more indicators are necessary to adequately measure the latent variable (Nunnally & Bernstein, 1994).

Models that include multidimensional indicators (observed variables that are predicted by two or more latent variables) can be more difficult to identify/estimate than standard CFA models with only unidimensional indicators (Kenny et al., 1998). This helps explain why more complex CFA models (e.g. multi-trait multi-method matrices; Campbell & Fiske, 1959) are notoriously difficult to estimate (Kenny & Kashy, 1992; Marsh & Grayson, 1995; Marsh & Hocevar, 1983). For CFA models that include correlated errors, in addition to ensuring that the number of knowns is equal to or greater than the number of unknowns, 'each latent variable needs two indicators that do not have correlated errors and every pair of latent variables needs at least one indicator that does not share correlated errors' (Kenny & Milan, 2012, p. 153).

In summary, standard CFA models (with unidimensional items and no correlated errors) are identified if the number of knowns is equal to or greater than the number of unknowns. However, the identification rules for models with correlated errors and models with multidimensional indicators are more complicated (Brown, 2015). For more details about the identification of such CFA models, see Kenny et al. (1998), who provide a thorough treatment of identification issues in CFA models.

Degrees of freedom for hybrid SEM

How many degrees of freedom does the hybrid SEM model in Figure 5.3 contain? There are seven observed variables: three indicators of academic persistence, three indicators of growth mindset and one indicator of academic achievement. Therefore, the number of knowns equals 7 * 8/2, or 28. How many parameters are freely estimated? Using the **fixed factor variance strategy**, we estimate six pattern coefficients (paths from latent variables to observed indicators of their respective factors), six measurement error variances, two disturbance variances and three structural paths (the paths among the conceptual variables of interest: growth mindset, academic persistence and academic achievement). Therefore, this model has 17 free parameters (unknowns). With 28 knowns and 17 freely estimated parameters, there are 11 *df* (28 – 17). All 11 *df* come from the measurement portion of the model. There are three structural variables: (1) growth mindset, (2) academic persistence and (3) academic achievement, and there are linkages among these three structural variables. Therefore, the structural portion of the model is just-identified. This means that any model misfit is due to misspecifying the measurement portion of the model. We return to this issue in Chapter 6, when we describe the model building process.

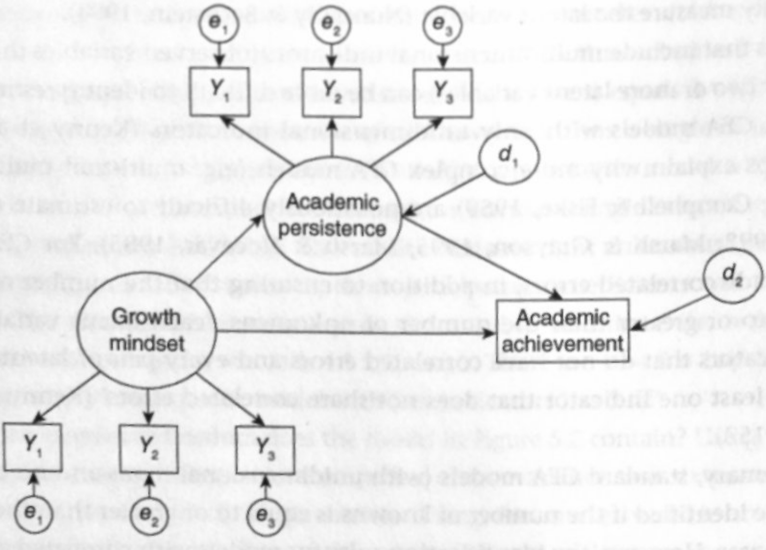

Figure 5.3 A hybrid structural equation model: A path model that includes latent variables

Equations for a measurement (CFA) model

Using the path diagram in Figure 5.2, we can represent the measurement model as a system of equations, as shown in Exhibit 5.1. The system of equations captures all the direct pathways among the variables but does not include (non-directional) correlations among variables. To start, let's write an equation for each of the endogenous variables in our model. The equation for each endogenous variable is analogous to a regression equation. The endogenous variable always appears on the left-hand side of the equation. Any (observed or latent) variable with a single-headed arrow leading to the endogenous variable appears on the right-hand side of the equation. Because standard structural equation models are linear models, the terms are additive.

Exhibit 5.1 Systems of equations corresponding to Figure 5.2

$\text{Algebra} = \lambda_1\text{Math} + \delta_1$

$\text{Geometry} = \lambda_2\text{Math} + \delta_2$

$\text{Measurement} = \lambda_3\text{Math} + \delta_3$

$\text{Comprehension} = \lambda_4\text{Reading} + \delta_4$

$\text{Vocabulary} = \lambda_5\text{Reading} + \delta_5$

$\text{Fluency} = \lambda_6\text{Reading} + \delta_6$

Introduction to systems of equations using path diagrams

Getting started with Wright's rules

Wright's tracing rules, developed in the 1910s and 1920s by biologist Sewall Wright (Heise, 1975), provide the basic principles of path analysis. Wright's **standardised tracing rules** provide the most intuitive method to generate the model-implied correlations from the standardised parameters for recursive path/structural equation models.

It is also possible to generate the model-implied variance–covariance matrix from the unstandardised parameter estimates using the **unstandardised tracing rules**. However, the standardised tracing rules are both more straightforward and more common than the unstandardised tracing rules, and the standardised path coefficients and correlations are generally easier to interpret. In Appendix 3, we provide the technical details that undergird our discussion of Wright's standardised tracing rules In Appendix 4, we discuss Wright's rules for generating the model-implied covariances using the unstandardised path coefficients for CFA and path models and we demonstrate the equivalence of using either unstandardised tracing rules or **covariance algebra** for generating model-implied covariances. These technical details provide a deeper understanding of the mathematical underpinnings of SEM.

Standardised tracing rules

Wright's tracing rules for standardised variables (Wright, 1918, 1934) are a set of rules for tracing the model which implies distinct correlations between two variables based on the structural relations between variables in a path diagram. The model-implied correlation matrix is essentially the standardised version of the model-implied covariance matrix. Using Wright's standardised tracing rules (Loehlin, 2004; Wright, 1918, 1934), we can determine the model-implied correlation between any two variables in a proper (recursive) path diagram using three simple rules:

1. No loops (you cannot pass through the same variable twice in one trace)
2. No going forward then backward within a given trace (but you can go backward then forward)
3. A maximum of one curved arrow per path

Rule 2 states that traces can go backward and then forward, but not forward and then backward, which may seem confusing and capricious at first glance. Why can we go backward and then forward but not forward and then backward? Conceptually, rule 2 accounts for linkages due to common causes, but not linkages due to common effects. We illustrate this idea with two simple three-variable systems of equations, depicted in Figures 5.4 and 5.5.

Figure 5.4 Tracing rule: You can go backward and then forward. This figure illustrates a linkage due to common causes (upstream variables)

Note. If *X* and *Y* are both predicted by *Z*, then they must be related to each other.

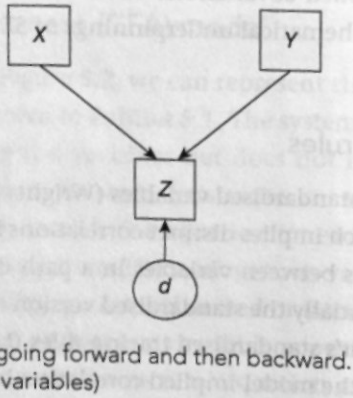

Figure 5.5 Tracing rule: No going forward and then backward. There is NO linkage due to common effects (downstream variables)

Note. *X* and *Y* can both predict *Z* and still be uncorrelated with each other. (They each predict different portions of the variance in *Z*.)

In Figure 5.4, *Z* predicts both *X* and *Y*. Because both *X* and *Y* share variance in common with *Z*, they must share variance in common with each other. Tracing backward from *Y* to *Z* and then forward from *Z* to *X* accounts for the variance that *Y* and *X* share, given that they share a common predictor (*Z*). In Figure 5.5, both *X* and *Y* predict *Z*. Two exogenous variables can predict the same endogenous variable without being related to each other. Similarly, two variables can be correlated with a third variable without being correlated with each other (Wright, 1934). In this case, *X* and *Y* are uncorrelated with each other (there is no curved arrow connecting *X* and *Y*), so each variable explains different portions of the variance in *Z*. The prohibition on tracing forward and then backward prevents counting linkages due to common effects when determining model-implied correlations.

Using these three rules, we can determine the model-implied (or expected) correlations among all the variables in the model. To do so, we sum all the **compound paths**, or traces, between two variables (i.e. direct and indirect as described in Chapter 4; Loehlin, 2004; Neale & Cardon, 1992). A compound path (trace) is a pathway connecting the two variables following the three rules above and is the product of all constituent paths (Loehlin, 2004). However, there may be many compound paths that connect the same set of two variables. To compute the model-implied correlation between two variables, first, take the product of all elements within each compound path. Then sum all the compound paths. In other words, the model-implied correlation involves multiplying each of the coefficients in a trace and summing over all possible traces (each trace is referred to as a compound path and we sum over the compound paths). Using these rules, we can compute the model-implied correlation between any two variables in a path diagram.

Example of the standardised tracing rule

Figure 5.6 is a path diagram with standardised path coefficients. Using the tracing rules, we can compute the model-implied correlations among all pairs of variables in the model (Table 5.1). The model-implied correlation between parental expectations and academic achievement is the sum of the compound paths (traces) connecting the two variables. What are all the potential traces from parental expectations to academic achievement? Using the tracing rule, there are three *traces* from parental expectations to academic achievement. The first is the direct effect of parental expectations on academic achievement: that path = .1. The second is the indirect effect of parental expectations on academic achievement through academic persistence: .3 * .5 = .15. The third is the indirect pathway through growth mindset and academic persistence: .2 * .4 * .5 = .04. The sum of these three compound paths (.1 + .15 + .04 = .29) is the model-implied correlation between parental expectations and academic achievement. In this case, it is also the total effect of parental expectations on academic achievement. The total effect and the model-implied correlation are identical when all traces involve only paths (i.e. none of the traces include correlations).

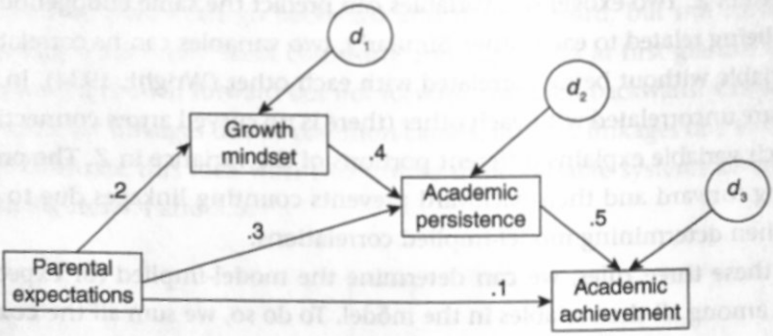

Figure 5.6 Using the tracing rules to compute model-implied correlations

Table 5.1 Model-implied correlations among the four observed variables in Figure 5.6

Variable	Parental Expectations	Growth Mindset	Academic Persistence	Academic Achievement
Parental Expectations	1.0			
Growth Mindset	.2	1.0		
Academic Persistence	.38	.46	1.0	
Academic Achievement	.29	.25	.54	1.0

What is the model-implied correlation of academic persistence and academic achievement? Again, we compute the model-implied correlation as the sum of the traces connecting the two variables. Using Wright's rules, there are three distinct traces that link academic persistence to academic achievement. The first is the most obvious: the direct effect of academic persistence on academic achievement has a standardised coefficient of .5. Rule 2 states that traces can go backward through arrowheads and then forward (but not forward and then backward). The second trace goes backward from academic persistence to growth mindset (b = .4), then backward from growth mindset to parental expectations (b = .2) and then forward from parental expectations to academic achievement. The product of these three paths is .4 * .2 * .1 = .008. The third trace goes backward from academic persistence to parental expectations (b = .3), and then forward from parental expectations to academic achievement (.1). The product of these two paths is .3 * .1 = .03. The model-implied correlation between academic persistence and academic achievement is the sum of these three traces (compound paths): .5 + .03 + .008 = .538 (which rounds to .54).

The model-implied correlation between growth mindset and academic persistence is .46. Why? There are two distinct traces (compound paths) linking growth mindset and persistence: a direct pathway (*b* = .40) and a trace that goes backward through parental expectations (*b* = .20) then forward from parental expectations to persistence (.30). That compound path is .06. The sum of the two compound paths results in a model-implied correlation of .46 (.40 + .06).

Finally, even though there is no direct effect of growth mindset on academic achievement, the model-implied correlation between growth mindset and academic achievement is not 0. Why? Following Wright's rules, there are actually three compound paths (traces) that link growth mindset and academic achievement. The first is the indirect effect of growth mindset on academic achievement via academic persistence, which is .4 * .5 = .20. The second compound path traces backward from growth mindset to parental expectations (*b* = .2) and then forward from parental expectations to academic achievement (*b* = .1). This compound path = .2 * .1 = .02. The third compound path traces backward from growth mindset to parental expectations (*b* = .2), then forward from parental expectations to academic persistence (.3), and then forward from academic persistence to academic achievement (.5), which equals .2 * .3 * .5 = .03. Summing these three compound paths (traces) (.20 + .02 + .03) yields the model-implied correlation, which is .25. In other words, even though the model constrains the direct effect of academic persistence on academic achievement to 0, the model-implied correlation between academic persistence on academic achievement is .25. Using the tracing rules above, confirm that the model-implied correlation between parental expectations and academic persistence is .38.

Standardised tracing rules for measurement models

We can apply the tracing rules to compute model-implied correlations in a standard CFA model. Each indicator is predicted by only one factor and there are no correlations among measurement errors, which greatly simplifies the tracing rules. Because there is only one compound path connecting any two variables, the model-implied correlation between the two variables of interest is simply the product of the paths and correlations connecting the two variables. If the factor model adequately explains the data, the correlation between any two indicators of the same factor should equal the product of the paths connecting them. The correlation between two indicators on two different factors should equal the product of the paths connecting each indicator to its respective factor multiplied by the correlation between the two factors.

Figure 5.7 contains standardised path coefficients and correlations for our CFA model. Using the standardised tracing rules, we can estimate the model-implied correlation between two observed variables in our model using the path (pattern) coefficients and correlation coefficients. For two observed variables within the same factor, the model-implied correlation is the product of the paths from the factor to each of the observed variables because we can go backward from one variable to the factor and then forward from the factor to the other variable. For example, the correlation between the algebra and geometry scores is the product of the standardised path coefficients for the two paths leading from the Math factor to these respective scores: .80 * .70 = .56. Likewise, the model-implied correlation between algebra and measurement is .8 * .6 = .48 and the model-implied correlation between geometry and measurement is .7 * .6 = .42.

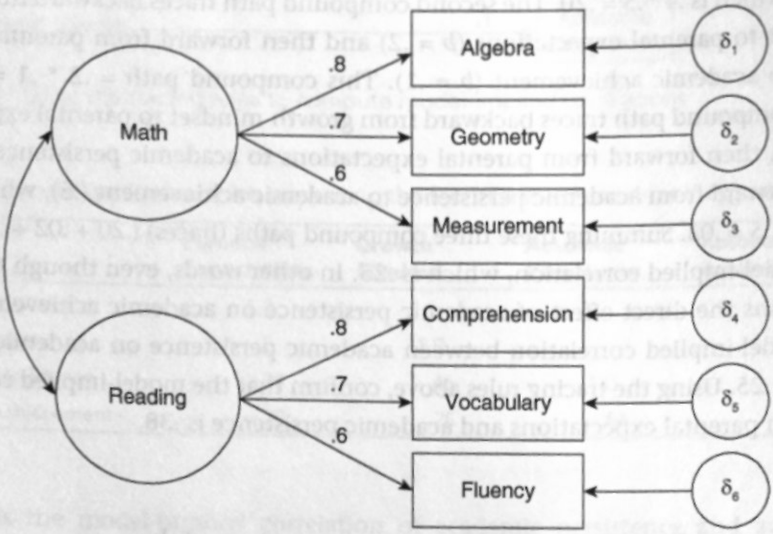

Figure 5.7 A two-factor CFA model with standardised path coefficients and correlations

To estimate the model-implied correlation between two observed variables from two different factors, we trace backward from observed variable 1 to its factor, then we trace through the correlation between the two factors (the curved arrow), and then we trace forward from factor 2 to observed variable 2. For example, the standardised path from the math factor to algebra scores is .80, the correlation between the math and reading factors is .50, and the standardised path coefficient from the reading factor to comprehension scores is .80. Therefore, the model-implied correlation between algebra scores and comprehension scores is .80 * .50 * .80, or .32. The model-implied correlation between geometry and fluency is .7 * .5 * .6, which equals .21. Be sure that you can compute all the model-implied correlations among the six observed variables in Table 5.2 using Figure 5.7.

Table 5.2 Model-implied correlations among the six observed variables in Figure 5.7

Variable	Algebra	Geometry	Measurement	Comprehension	Vocabulary	Fluency
Algebra	1.0					
Geometry	.56	1.0				
Measurement	.48	.42	1.0			
Comprehension	.32	.28	.24	1.0		
Vocabulary	.28	.245	.21	.56	1.0	
Fluency	.24	.21	.18	.48	.42	1.0

Using the tracing rule, we can also compute the model-implied correlations between the factors and the observed variables. In factor analysis, these are generally referred to as the *structure coefficients*. The model-implied correlation between the math factor and comprehension scores is the product of the correlation between the math factor and the reading factor (.50) and the standardised path from the reading factor to comprehension scores (.80) = .80 * .50 = .40. So even though the direct path from the math factor comprehension scores is 0, the model-implied correlation between the math factor and reading comprehension scores is .40 (and the model-implied correlation between algebra and reading comprehension scores is .32).

In a standard CFA model, because the measurement error terms (δ's) are independent of each other and of the factors, they do not contribute to the estimation of the model-implied correlations among the observed variables in the standard CFA model. Forcing the measurement errors to be uncorrelated with each other and with the factors specifies that the error variances in the observed indicators (the residual observed score variance that is not explained by their respective factors) are independent (uncorrelated with each other). Allowing the errors of two variables to correlate with each other indicates that factor structure does not adequately capture the correlation between the two variables: the two variables are either more or less correlated with each other than would be predicted by the CFA model. Usually, the correlation between the error variances is positive, indicating that the observed correlation of the two observed variables is higher than would be explained by the factor structure: the two indicators share something in common that is not explained by the factor(s). Generally, correlated errors are indicative of unmodelled multidimensionality. Correlated errors may signal the presence of method effects (e.g. mode of data collection, wording effects) or substantive similarities among variables that are not fully captured by the specified factor structure. Therefore, adding correlated errors to a model should be substantively motivated and conceptually defensible. Although we recommend correlating errors sparingly, sometimes it is appropriate or even necessary to correlate errors. For example, when conducting longitudinal CFAs, it is common practice to allow the measurement error terms of the same indicator to correlate across time. Why? If the exact same measure is administered at multiple time points, it seems quite plausible that the unique variance in that measure would correlate across time, even after controlling for the latent variable at each time point. Correlating the errors allows the unexplained variance in an indicator at one time point to be related to the unexplained variance in the same indicator, measured at a different time point.

What happens when we add correlated errors to a CFA/measurement model? Imagine that we add a correlation between the measurement errors of the comprehension and vocabulary scores (Figure 5.8). Both indicators load on the same latent variable (reading) and they are both unidimensional (only one factor predicts each of the indicators). Therefore, the two indicators (a) load on the same factor, (b) have no cross-loadings with any other factors and (c) do not have correlated errors with any other indicators. In this restrictive scenario, adding a correlation between the two error terms perfectly reproduces the correlation between the two indicators (observed variables). Why? The correlation between the residuals can take on any positive or negative value (between –1 and 1) that best reproduces the correlation between the two indicators. Regardless of the parameter estimates of the two direct paths (factor loadings), it is possible to specify an error correlation that perfectly reproduces the correlation between the observed variables. Under such conditions, adding a correlation between two measurement errors of two indicators of the same factor eliminates the correlation between those two variables as a source of

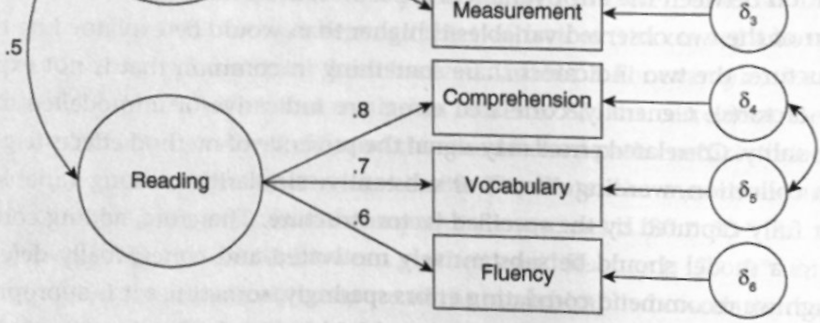

Figure 5.8 A two-factor CFA model with standardised parameters and a correlated error

information that helps us to understand the factor structure of that latent construct. (However, if the indicators load on more than one factor, including cross-loadings, or if either of the indicators shares a correlated error with another indicator as well, then correlating their measurement errors does not exactly reproduce the correlation between the indicators.)

In Chapter 6, we outline a sequence of model building steps. Then, using all that we have learned, we fit and interpret a latent variable (hybrid) structural equation model.

Chapter Summary

- Identification involves demonstrating 'that the unknown parameters are functions only of the identified parameters and that these functions lead to unique solutions' (Bollen, 1989, p. 88).
- The number of knowns places an upper limit on the number of freely estimated parameters in the model (the unknowns). These unknowns are the parameters that we wish to estimate.
- A just-identified model contains as many knowns as unknowns, so the parameter estimates can always perfectly reproduce the variance–covariance matrix.
- An overidentified model uses a smaller number of parameters to estimate all elements of the variance–covariance matrix, resulting in some discrepancy between the available variance–covariance matrix and the parameters to be estimated.
- If the specified model requires estimating more parameters than there are unique pieces of information in the variance–covariance matrix, the model has negative degrees of freedom and is underidentified. It is not possible to solve the set of structural equations for underidentified models because there are more unknowns than knowns.
- In standard, single-group CFA, two scaling strategies result in statistically equivalent models: the fixed factor variable and marker variable strategies.

Further Reading

Kenny, D. A. (2004). *Correlation and causality* (Rev. ed.). Wiley-Interscience. http://davidakenny.net/doc/cc_v1.pdf

Kenny's classic book is out of print, but the pdf is available at the address above. This book explains identification rules and covariance algebra in great detail.

Kenny, D. A., & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 145–163). Guilford Press.

As the title states, this book chapter by Kenny and Milan provides a non-technical discussion of the technical issue of identification. The chapter provides an overview and rules of thumb for identification. Furthermore, the chapter discusses identification for path analysis models without feedback (recursive), with feedback (non-recursive) and with omitted variables as well as latent variable models and latent growth models.

6

BUILDING STRUCTURAL EQUATION MODELS

Chapter Overview